


# Reconstructing Voice Identity from Noninvasive Auditory Cortex Recordings

Reviewed Preprint


v1 • July 15, 2024

Not revised

Charly Lamothe , Etienne Thoret, Régis Trapeau, Bruno L Giordano, Julien Sein, Sylvain Takerkart, Stéphane Ayache, Thierry Artières , Pascal Belin 

La Timone Neuroscience Institute UMR 7289, CNRS, Aix-Marseille University, Marseille, France • Laboratoire d'Informatique et Systèmes UMR 7020, CNRS, Aix-Marseille University, Marseille, France • Perception, Representation, Image, Sound, Music UMR 7061, CNRS, Marseille, France • Institute of Language Communication & the Brain, Marseille • Centre IRM-INT@CERIMED, Marseille, France • École Centrale de Marseille, Marseille, France

 [https://en.wikipedia.org/wiki/Open\\_access](https://en.wikipedia.org/wiki/Open_access)

 Copyright information

## Abstract

The cerebral processing of voice information is known to engage, in human as well as non-human primates, “temporal voice areas” (TVAs) that respond preferentially to conspecific vocalizations. However, how voice information is represented by neuronal populations in these areas, particularly speaker identity information, remains poorly understood. Here, we used a deep neural network (DNN) to generate a high-level, small-dimension representational space for voice identity—the ‘voice latent space’ (VLS)—and examined its linear relation with cerebral activity via encoding, representational similarity, and decoding analyses. We find that the VLS maps onto fMRI measures of cerebral activity in response to tens of thousands of voice stimuli from hundreds of different speaker identities and better accounts for the representational geometry for speaker identity in the TVAs than in A1. Moreover, the VLS allowed TVA-based reconstructions of voice stimuli that preserved essential aspects of speaker identity as assessed by both machine classifiers and human listeners. These results indicate that the DNN-derived VLS provides high-level representations of voice identity information in the TVAs.

### eLife assessment

This study used deep neural networks (DNN) to reconstruct voice information (viz., speaker identity), from fMRI responses in the auditory cortex and temporal voice areas, and assessed the representational content in these areas with decoding. A DNN-derived feature space approximated the neural representation of speaker identity-related information. While some of the neural decoding results are **valuable**, the overall evidence for general representational and computational principles is **incomplete** as the results rely on a very specific model architecture.

<https://doi.org/10.7554/eLife.98047.1.sa3>

## Introduction

The human voice carries speech, but is also an “auditory face” that carries much valuable information on the stable physical characteristics of the speaker (hereafter, ‘identity-related’; [Belin et al., 2004](#) , [2011](#) ). The ability of listeners to extract identity-related information in voice such as gender, age, or unique identity even in brief stimuli plays a crucial role in our social interactions, yet its neural bases remain poorly understood compared to those of speech processing. Studies over the past two decades have clearly established via complementary neuroimaging techniques that the cerebral processing of voice information involves a set of temporal voice areas (TVAs) in secondary auditory cortical regions of the human (fMRI: [Belin et al., 2000](#) , [von Kriegstein et al., 2004](#), [Pernet et al., 2015](#) ; EEG, MEG: [Charest et al., 2009](#) , [Capilla et al., 2013](#) , [Barbero et al., 2021](#) ; Electrophysiology: [Rupp et al., 2022](#) , [Zhang et al., 2021](#) ) as well as macaque brain ([Petkov et al., 2008](#) ; [Bodin et al., 2021](#) ). The TVAs respond more strongly to sounds of voice – with or without speech ([Pernet et al., 2015](#) ; [Rupp et al., 2022](#) ; [Trapeau et al., 2022](#) )—and categorize voice apart from other sounds ([Bodin et al., 2021](#) ) but the nature of the information encoded at these stages of cortical processing, especially with respect to speaker identity-related information, remains largely unknown ([Blank et al., 2014](#) ; [Belin et al., 2018](#) ).

In recent years, deep neural networks (DNNs) have emerged as a powerful tool for representing complex visual data, such as images ([LeCun et al., 2015](#) ) or videos ([Liu et al., 2020](#) ). In the auditory domain, DNNs have been shown to provide valuable representations—so-called feature or latent spaces—for modeling the cerebral processing of sound (brain encoding) (speech: [Kell et al., 2018](#) ; [Millet et al., 2022](#) ; [Tuckute & Feather, 2023](#); semantic content: [Caucheteux et al., 2022](#) ; [Caucheteux & King, 2022](#) ; [Caucheteux et al., 2023](#) ; [Giordano et al., 2023](#) ; music: [Güçlü et al., 2016](#) ), or reconstructing the stimuli listened by a participant (brain decoding) ([Akbari et al., 2019](#) ). They have not yet been used to explain cerebral representations of identity-related information due in part to the focus on speech information ([von Kriegstein et al., 2003](#) ).

Here, we addressed this challenge by training a ‘Variational autoencoder’ (VAE; [Kingma et Welling, 2014](#)) DNN to reconstruct voice spectrograms from 182,000 250-ms voice samples from 405 different speaker identities in 8 different languages from the CommonVoice database ([Ardila et al., 2020](#) ). Brief (250 ms) samples were used to emphasize speaker identity-related information in voice, already available after a few hundred milliseconds ([Schweinberger et al., 1997](#) ; [Lavan, 2023](#) ), over linguistic information unfolding over longer periods (word, >350 ms; [McAllister et al., 1994](#) ). While a quarter of a second is admittedly short compared to standards of, e.g., computational speaker identification that typically uses 2–3 s samples, this short duration is sufficient to allow near-perfect gender classification and performance levels well above chance for speaker discrimination (**Fig. 5d** , red dotted line). This brief duration allowed the presentation of many more stimuli to our participants in the scanner while preserving acceptable behavioral and classifier performance levels.

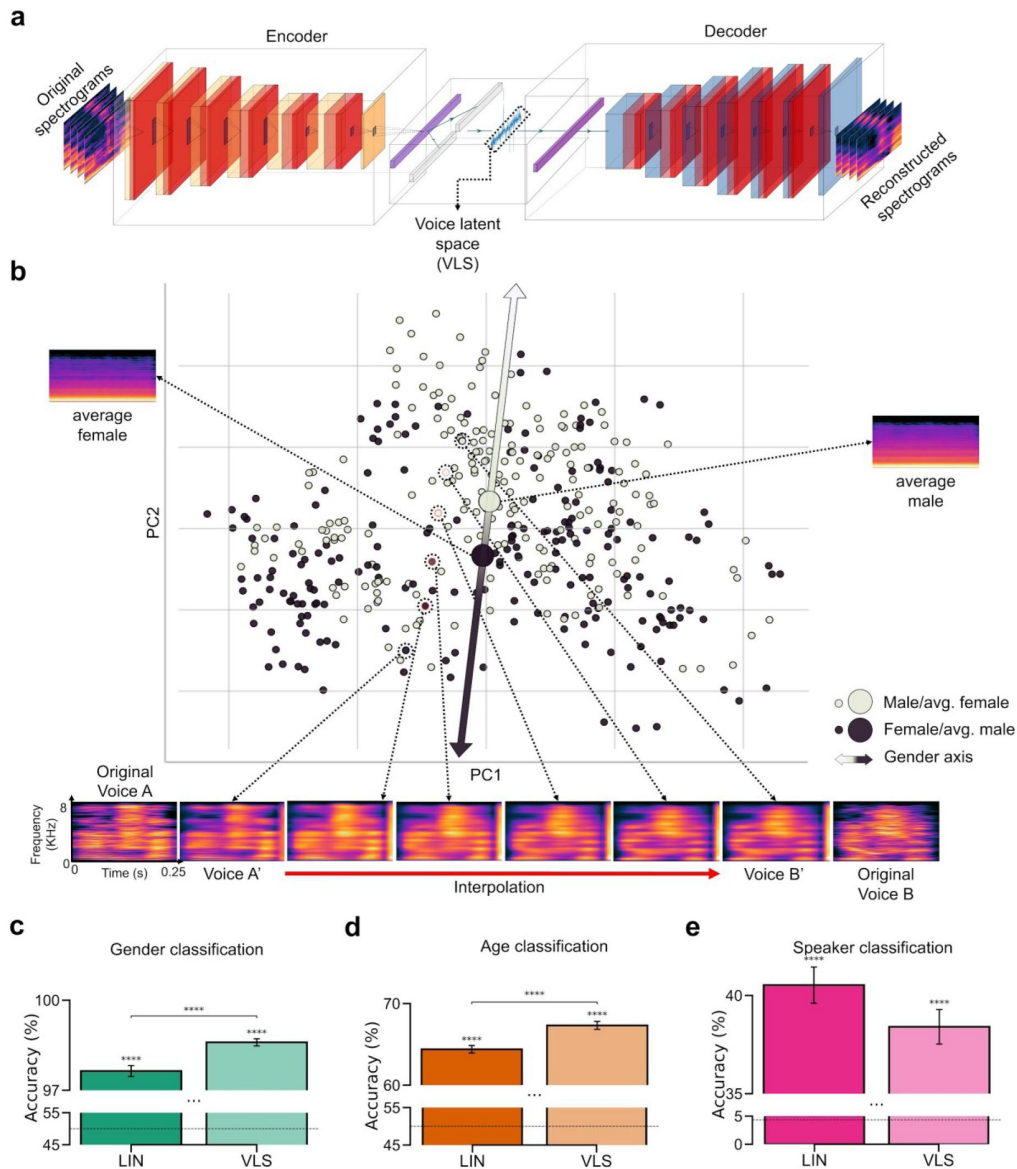
State-of-the-art studies have primarily relied on task-optimized neural networks (i.e., DNN trained using supervised learning to classify a category from the input) to study sensory cortex processes ([Yamins & DiCarlo, 2016](#) ; [Schrimpf et al., 2018](#) ). They can reach high accuracies in brain encoding ([Khaligh-Razavi & Kriegeskorte, 2014](#) ; [Schrimpf et al., 2018](#) ; [Han et al., 2019](#) ). However, there is increasing evidence that unsupervised learning, such as that used for the VAE, also provides plausible computational models for investigating brain processing ([Higgins et al., 2021](#) ; [Zhuang et al., 2021](#); [Millet et al., 2022](#) ; [Orhan et al., 2022](#)). Thus, the VAE-derived VLS, exploited within encoding, representational similarity, and decoding frameworks, offers a potentially promising tool for investigating the representations of voice stimuli in the secondary auditory cortex ([Naselaris et al., 2011](#) ). Autoencoders learn to compress stimuli with high

dimensionality into a lower-dimensional space that nonetheless allows reconstruction of the original stimuli via an inverse transformation learned by the second part of the network called the decoder. **Fig. 1a** shows the architecture of the VAE, with its encoder that reduces an input spectrogram to a highly compressed, 128-dimension *voice latent space* (VLS) representation and its decoder that reconstructs the spectrogram from this VLS representation. We selected this latent space size as it was the first value that produced satisfactory reconstructions. Points in the VLS correspond to voice samples with different identities and phonetic content. A line segment in the VLS contains points corresponding to perceptual interpolations between its two extremities (**Fig. 1b**; Supplementary Audio 1). VLS coordinates of samples presented to the participants averaged by speaker identity suggest that a major organizational dimension of the latent space is voice gender (**Fig. 1b**) (colored by age or language in Supplementary Figure 1).

In order to test whether VLS accounts well for cerebral activity in response to voice stimuli, we scanned three healthy volunteers using fMRI to measure an indirect index of their cerebral activity across 10+ hours of scanning each in response to ~12,000 of the voice samples, denoted *BrainVoice* in the following (different from the ones used to train the DNN). The small number of participants does not allow for generalization at the general population level as in standard fMRI studies. However, it allows testing for replicability as in comparable studies involving 10+ hours of scanning per participant (VanRullen & Reddy, 2019). Different stimulus sets were used across participants to provide a stringent test of replicability based on subject-level analyses. Stimuli consisted of randomly spliced 250-ms excerpts of speech samples from the CommonVoice database (Ardila et al., 2020) by 119 speakers in 8 languages. For assessing generalization performances of decoding models and brain-based reconstruction, six test stimuli were repeated more often (60 times) for each participant to provide robust estimates of their induced cerebral activity (see Methods). We first modeled these responses to voice using a general linear model (GLM) (Friston et al., 1994) with several nuisance regressors as an initial denoising step (Supplementary Figure 3), then used a second GLM modeling cerebral responses to the different speaker identities (Supplementary Figure 2a), resulting in one voxel activity map per speaker (Supplementary Figure 2b). We independently localized in each participant several regions of interest (ROIs) on which subsequent analyses were focused: the anterior, middle and posterior TVAs in each hemisphere (individually localized via an independent ‘voice localizer scan’ and MNI coordinates provided in Pernet et al., 2015; Supplementary Figure 2c) as well as primary auditory cortex (A1) (using a probabilistic map in MNI space (Penhune et al., 1996; Supplementary Figure 2d).

We first asked how the VLS could account for the brain responses to speaker identities (encoding) measured in A1 and the TVAs, in comparison with a linear autoencoder’s latent space (LIN). This approach was chosen to compare a representation learned linearly under similar conditions (same input data, learning algorithm, reconstruction objective and latent space size) with the VLS, which has non-linear transformations and a regularized latent space. For this, we used a general linear model (GLM) of fMRI responses to the speaker identities, resulting in one voxel activity map per speaker (Supplementary Figure 2). Then, we computed the average VLS coordinates of the fMRI voice stimuli for each speaker identity, which may be seen as a speaker representation in the VLS (see *Identity-based and stimulus-based representations* section). Next we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker VLS coordinates. As VAE achieves compression through a series of nonlinear transformations (Wetzel, 2017), we choose to contrast its results with a linear autoencoder’s latent space. This method has previously been applied to fMRI-based image reconstructions (Cowen et al., 2014; VanRullen & Reddy, 2019; Mozafari et al., 2020).

The extent to which the VLS allows linearly predicting the fMRI recordings does not provide insight into the representational geometries, i.e., the differences between the patterns of cerebral activity for speaker identity. We addressed this question by using representational similarity analysis (RSA; Kriegeskorte et al., 2008) to test which model better accounts for the representational geometry for voice identities in the auditory cortex. Using RSA as a model



**Fig. 1.**

### DNN-derived Voice Latent Space (VLS).

**a**, Variational autoencoder (VAE) Architecture. Two networks learned complementary tasks. An encoder was trained using 182K voice samples to compress their spectrogram into a 128-dimension representation, the voice latent space (VLS), while a decoder learned the reverse mapping. The network was trained end-to-end by minimizing the difference between the original and reconstructed spectrograms. **b**, Distribution of the 405 speaker identities along the first 2 principal components of the VLS coordinates from all sounds, averaged by speaker identity. Each disk represents a speaker's identity colored by gender. PC2 largely maps onto voice gender (ANOVAs on the first two components: PC1:  $F(1, 405)=0.10$ ,  $p=.74$ ; PC2:  $F(1, 405)=11.00$ ,  $p<.001$ ). Large disks represent the average of all male (black) or female (gray) speaker coordinates, with their associated reconstructed spectrograms (note the flat fundamental frequency ( $f_0$ ) and formant frequencies contours caused by averaging). The bottom of the spectrograms illustrates an interpolation between stimuli of two different speaker identities: spectrograms at the extremes correspond to two original stimuli (A, B) and their VLS-reconstructed spectrograms (A', B'). Intermediary spectrograms were reconstructed from linearly interpolated coordinates between those two points in the VLS (red line) (cf. Supplementary Audio 1). **c, d, e**, Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (young/adult, chance level: 50%), or identity (119 identities, chance level: 0.84%) based on VLS or LIN coordinates. Error bars indicate the standard error of the mean (s.e.m) across 100 random classifier initializations. All  $p<1e-10$ . The horizontal black dashed lines indicate chance levels. \*\*\*\*:  $p<0.0001$ .

comparison framework is relevant to examining the brain-model relationship from complementary angles (Diedrichsen & Kriegeskorte, 2017 [DOI](#); Giordano et al., 2023 [DOI](#); Tuckute & Feather, 2023). We built speaker x speaker representational dissimilarity matrices (RDMs) capturing pairwise differences in cerebral activity or model predictions between all pairs of speakers; then, we examined how well the LIN and VLS-derived RDMs correlated with the cerebral RDMs from A1 and the TVAs.

A robust test of the adequacy of models of brain activity, and a long-standing goal in computational neurosciences, is the reconstruction of a stimulus presented to a participant from the evoked brain responses. While reconstruction of visual stimuli (images, videos) from cerebral activity has been performed by a number of groups (VanRullen & Reddy, 2019 [DOI](#); Mozafari et al., 2020 [DOI](#); Le et al., 2022 [DOI](#); Gaziv et al., 2022 [DOI](#); Dado et al., 2022 [DOI](#); Chen et al., 2023 [DOI](#)), validating the DNN-derived representational spaces, comparable work in the auditory domain is scarce, almost exclusively concentrated on linguistic information (Santoro et al., 2017 [DOI](#)). Akbari et al. (2019) [DOI](#) used a DNN to reconstruct speech stimuli based on ECoG recording of auditory cortex activity, an invasive method compared to techniques like fMRI. They obtained a good phonetic recognition rate but chance-level gender categorization performance from reconstructed spectrograms and no evaluation of speaker identity discrimination.

Here, we built on the linear relationship uncovered in our encoding analysis between the VLS and the fMRI recordings to invert it and try to predict VLS coordinates from the recorded fMRI data; then, using the decoder, we reconstructed the spectrograms of stimuli presented to the participants (Wu et al., 2006 [DOI](#); Naselaris et al., 2011 [DOI](#)). The voice identity information available in the reconstructed stimuli was finally assessed by human listeners using both machine learning classifiers and behavioral tasks (Fig. 4 [DOI](#)).

## Results

### Voice Information in the Voice Latent Space (VLS)

In order to probe the informational content of the VLS, linear classifiers were trained to categorize the voice stimuli from 405 speakers by gender (2 classes), age (2 classes) or identity (119 classes, cf Methods) based on VLS coordinates, or their LIN features as control (Fig. 1c,d,e [DOI](#); we aggregated the stimuli from the 3 participants; for each model computed the latent space of each stimulus and averaged the latent spaces by speaker identity, leading to 405 128-dimensional vectors. We then trained linear classifiers using a 5-fold cross-validation scheme, see *Characterization of the autoencoder latent space*). The mean of the distribution of accuracies obtained for 100 random classifier initializations (as to account for variance; Bouthillier et al., 2021 [DOI](#)) was significantly above chance level (all  $p$ s <  $1e-10$ ) for all classifications (LIN: gender (mean accuracy  $\pm$  s.d.) =  $97.64 \pm 1.77\%$ ,  $t(99)=266.94$ ; age:  $64.39 \pm 4.54\%$ ,  $t(99)=31.53$ ; identity:  $40.52 \pm 9.14\%$ ,  $t(99)=39.37$ ; VLS: gender:  $98.59 \pm 1.19\%$ ,  $t(99)=406.47$ ; age:  $67.31 \pm 4.86\%$ ,  $t(99)=35.41$ ; identity:  $38.40 \pm 8.75\%$ ,  $t(99)=38.73$ ). We then evaluated the difference in performance at preserving identity-related information between the VLS and LIN via one-way ANOVAs. Results showed a significant effect of Feature (LIN/VLS) in categories (all  $F$ s(1, 198) > 225.15, all  $p$ s < .0001) but not in identity. Post-hoc paired t-tests showed that the VLS was better than the LIN at encoding information related to voice identity, as evidenced by a significant difference in means for gender ( $t(99)=-6.11$ ,  $p<.0001$ ), age ( $t(99)=-6.10$ ,  $p<.0001$ ) but not for identity classifications ( $t(99)=1.71$ ).

Thus, despite its low number of dimensions (each input spectrogram has  $401 \times 21 = 8421$  parameters and is summarized in the VLS by a mere 128 dimensions), the VLS appears to meaningfully represent the different sources of voice information perceptually available in the vocal stimuli. This representational space, therefore, constitutes a relevant candidate for linearly modeling voice stimulus representations by the brain.



## Brain Encoding

We used a linear voxel-based encoding model to test whether VLS linearly maps onto cerebral responses to speaker identities measured with fMRI in the different ROIs. A regularized linear regression model (cf. Methods) was trained on a subset of the data (5-fold cross-validation scheme) to predict the voxel maps for each speaker identity. For each fold, the trained model was tested on the held-out speaker identities (**Fig. 2a**). The model's performance was assessed for each ROI using the Pearson correlation score between each voxel's actual and predicted responses (Schrimpf et al., 2021). Similar predictions were tested with features derived from LIN (cf. Methods). **Fig. 2b** shows the distribution of correlation coefficients obtained for each of the ROIs for the 2 sets of features across voxels, hemispheres, and participants.

One-sample t-tests showed that the means of Fisher z-transformed coefficients for both LIN features and VLS were significantly higher than zero (LIN: A1  $t(197)=7.25$ ,  $p<.0001$ , pTVA  $t(175)=4.49$ ,  $p<.0001$ , mTVA  $t(164)=9.12$ ,  $p<.0001$  and aTVA  $t(147)=6.81$ ,  $p<.0001$ ; VLS: A1  $t(197)=4.76$ ,  $p<.0001$ , mTVA  $t(164)=10.12$ ,  $p<.0001$  and aTVA  $t(147)=5.52$ ,  $p<.0001$  but not pTVA  $t(175)=-1.60$ ) (Supplementary Tables 2-3).

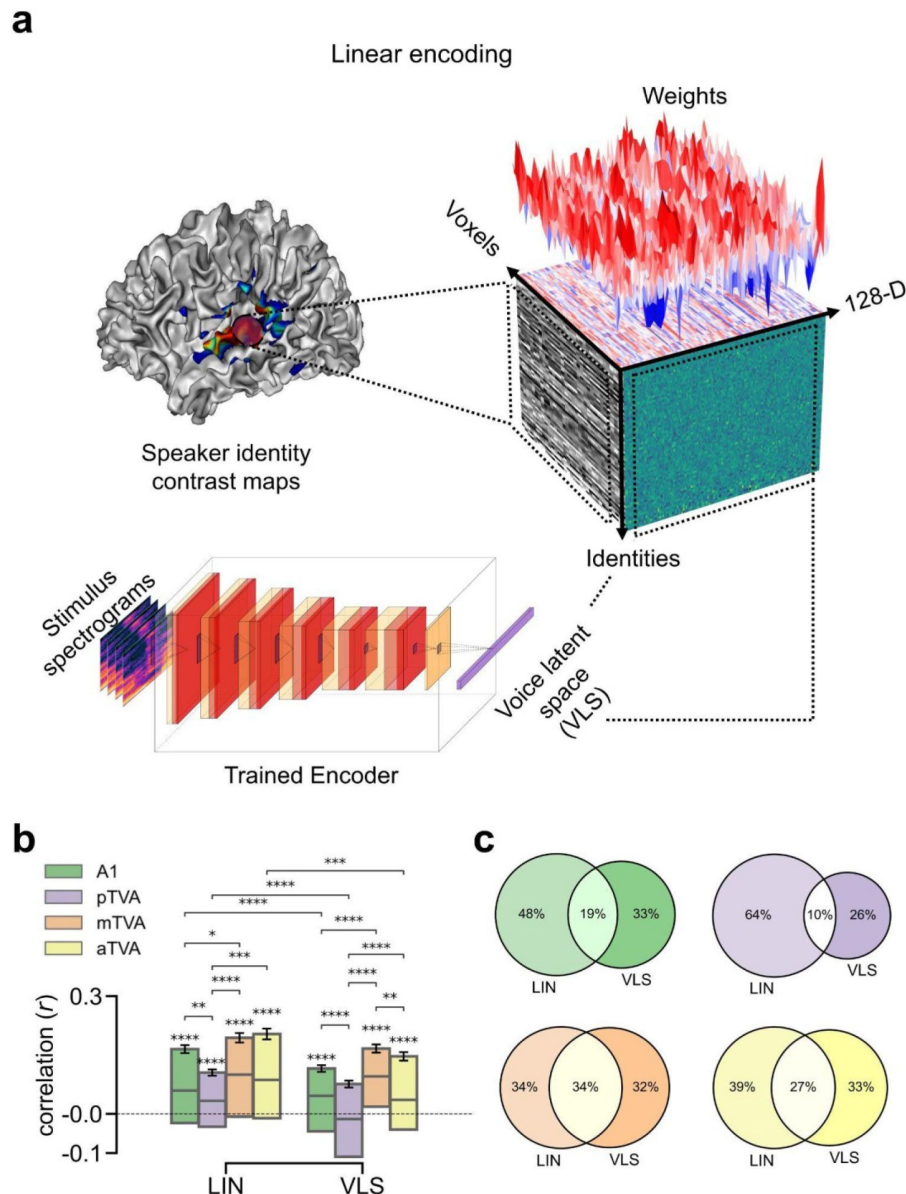
A mixed ANOVA performed on the Fisher z-transformed coefficients with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors showed a significant effect of Feature ( $F(3, 683)=56.65$ ,  $p<.0001$ ), a significant effect of ROI ( $F(3, 683)=18.50$ ,  $p<.0001$ ), and a moderate interaction Feature  $\times$  ROI ( $F(3, 683)=5.25$ ,  $p<.01$ ). Post-hoc comparisons revealed that the mean of correlation coefficients was higher for LIN than for VLS in A1 ( $t(197)=4.02$ ,  $p<.0001$ , pTVA  $t(175)=6.64$ ,  $p<.0001$ , aTVA  $t(147)=3.78$ ,  $p<.001$ ) but not in mTVA ( $t(164)=0.58$ ) (Supplementary Table 4); and that the voxel patterns are better predicted in mTVA than in A1 for both models (LIN:  $t(361)=2.36$ ,  $p<.05$ ; VLS:  $t(361)=4.91$ ,  $p<.0001$ ) (Supplementary Table 5). However, inspecting the distribution of model-voxel correlations, we found that both models account for different parts of the voice identity responses and differ across ROIs (**Fig. 2c**).

## Representational Similarity Analysis

For RSA, we built speaker  $\times$  speaker representational dissimilarity matrices (RDMs), capturing for each ROI the dissimilarity in voxel space between each pair of speaker voxel maps ('brain RDMs'; cf. Methods) using Pearson's correlation (Walther et al., 2016). We compared these four bilateral brain RDMs (A1, aTVA, mTVA, pTVA) to two 'model RDMs' capturing speaker pairwise feature differences predicted by LIN and the VLS (**Fig. 3a**) built using cosine distance (Xing et al., 2015; Bhattacharya et al., 2017; Wang et al., 2018). **Fig. 3b** shows for each ROI the Spearman correlation coefficients between the brain RDMs and the two model RDMs for each participant and hemisphere (Kriegeskorte et al., 2008; **Fig. 3c** for an example of brain-model correlation).

These brain-model correlation coefficients were compared to zero using a 'maximum statistics' approach based on random permutations of the model RDMs' rows and columns (Maris & Oostenveld, 2007; cf. Methods; **Fig. 3b**). For the LIN model, only one brain-model RDM correlation was significantly different from zero (one-tailed test): in mTVA, right hemisphere in S3 ( $p=.0500$ ). For the VLS model, in contrast, 5 significant brain-model RDM correlations were observed in all four ROIs: in A1, right hemisphere in S3 ( $p=.0142$ ); pTVA: right hemisphere in S3 ( $p=.0160$ ); mTVA: left hemisphere in S3 ( $p=.007$ ); aTVA: left hemispheres in S1 ( $p=.0417$ ) and S3 ( $p=.0001$ ) (Supplementary Table 6).

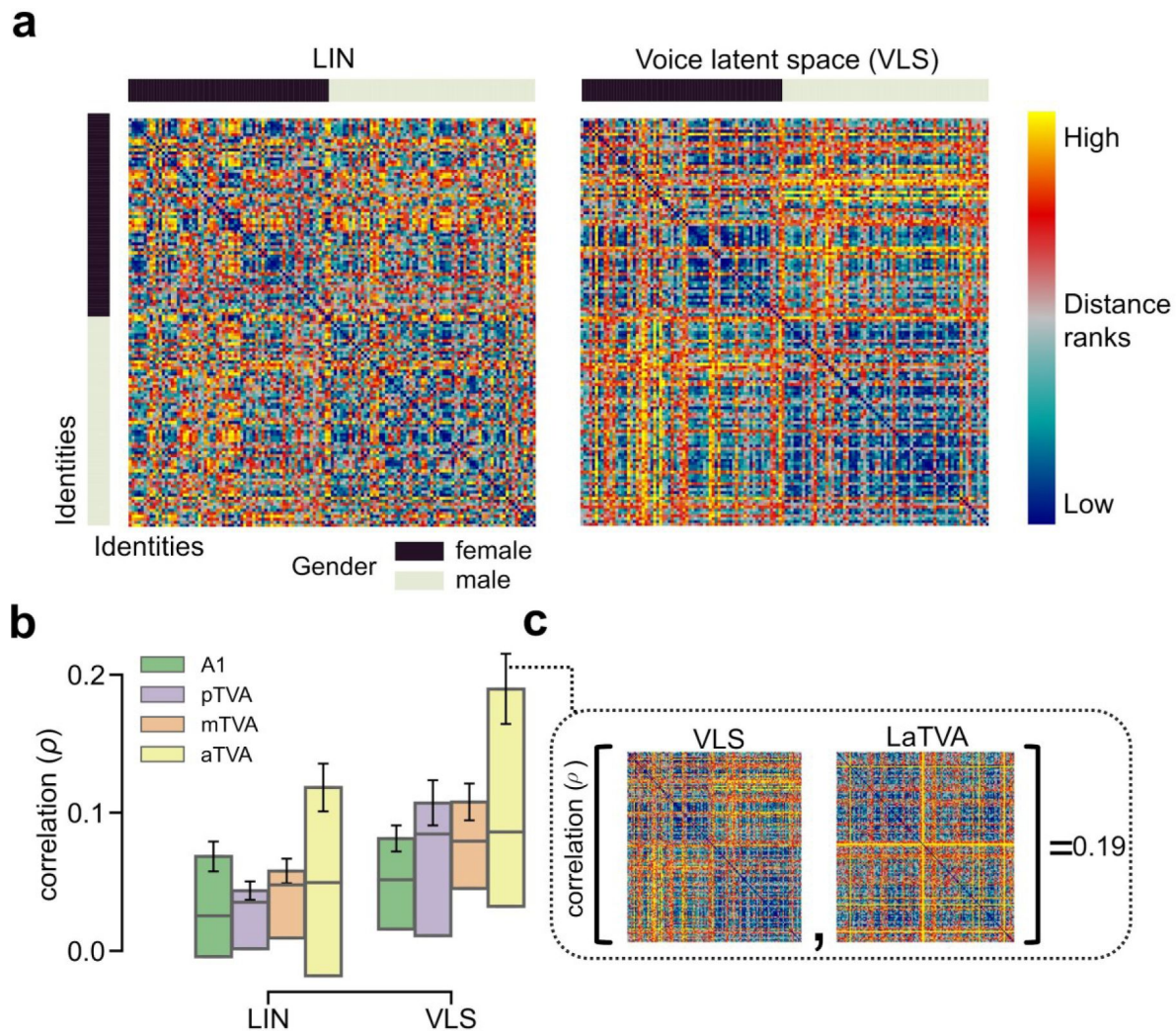
A two-way repeated-measures ANOVA with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on the Fisher z-transformed correlation coefficients showed a tendency towards a significant effect of Feature ( $F(1, 2)=22.53$ ,  $p=.04$ ), and no ROI ( $F(3, 6)=1.79$ ,  $p=.30$ ) or interaction effects ( $F(3, 6)=1.94$ ,  $p=.22$ ). We compared the correlation coefficients between the VLS



**Fig. 2.**

### Predicting brain activity from the VLS.

**a**, Linear brain activity prediction from VLS for ~135 speaker identities in the different ROIs. We first fit a GLM to predict the BOLD responses to each voice speaker identity. Then, using the trained encoder, we computed the average VLS coordinates of the voice stimuli presented to the participants based on speaker identity. Finally, we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker VLS coordinates. The cube illustrates the linear relationship between the fMRI responses to speaker identity and the VLS coordinates. The left face of the cube represents the activity of the voxels for each speaker's identity, with each line corresponding to one speaker. The right face displays the VLS coordinates for each speaker's identity. The cube's top face shows the encoding model's weight vectors. **b**, Encoding results. For each region of interest, the model's performance was assessed using the Pearson correlation score between the true and the predicted responses of each voxel on the held-out speaker identities. Pearson's correlation coefficients were computed for each voxel on the speakers' axis and then averaged across hemispheres and participants. Similar predictions were tested with the LIN features. Error bars indicate the standard error of the mean (s.e.m) across voxels. \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ; \*\*\*\* $p < 0.0001$ . **c**, Venn diagrams of the number of voxels in each ROI with the LIN, the VLS, or both models. For each ROI and each voxel, we checked whether the test correlation was higher than the median of all participant correlations (intersection circle), and if not, which model (LIN or VLS) yielded the highest correlation (left or right circles).



**Fig. 3.**

**The VLS better explains representational geometry for voice identities in the TVAs than the linear model.**

**a**, Representational dissimilarity matrices (RDMs) of pairwise speaker dissimilarities for ~135 identities (arranged by gender, cf. sidebars), according to LIN and VLS. **b**, Spearman correlation coefficients between the brain RDMs for A1, the 3 TVAs, and the 2 model RDMs. Error bars indicate the standard error of the mean (s.e.m) across brain-model correlations. **c**, Example of brain-model RDM correlation in the TVAs. The VLS RDM and the brain RDM yielding one of the highest correlations (LaTVA) are shown in the insert.



and LIN models within participants and hemispheres using one-tailed tests, based on the a priori hypothesis that the VLS models would exhibit greater brain-model correlations than the LIN models (cf. Methods). The results revealed two significant differences in one of the three participants, both favoring the VLS model (S3: right pTVA,  $p=.0366$ ; left aTVA,  $p=.00175$ ) (Supplementary Table 7).

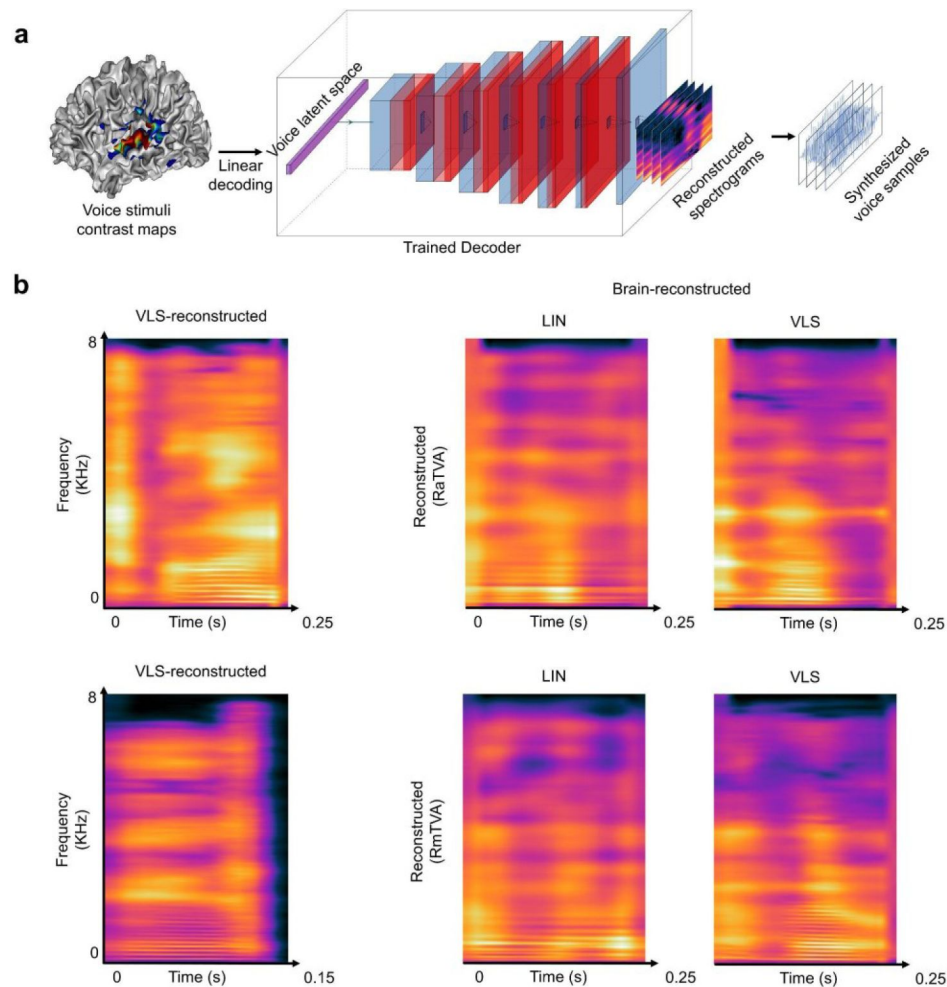
## Decoding and Reconstruction

We finally inverted the brain-VLS relationship to predict linearly VLS coordinates based on fMRI measurements (**Fig. 4a**; see ‘Brain decoding’ in Methods) and reconstructed via the trained decoder the spectrograms of 18 Test Stimuli (3 participants x 6 stimuli per participant; see **Fig. 4b**, and Supplementary Audio 2; audio estimated from spectrogram through phase reconstruction).

We first assessed the nature of the reconstructed stimuli by using a DNN trained to categorize natural audio events (Howard et al., 2017): all reconstructed versions of the 18 Test Stimuli were categorized as ‘speech’ (1 class out of 521 - no ‘voice’ classes). To evaluate the preservation of voice identity information in the reconstructed voices, pre-trained linear classifiers were used to classify the speaker gender (2 classes), age (2 classes), and identity (17 classes; one identity was shared across participants) of the 18 reconstructed Test Stimuli. The mean of the accuracy distribution obtained across random classifier initializations (20 per ROI) used on the stimuli reconstructed from the induced brain activity was significantly above chance level for gender (LIN: pTVA (mean accuracy  $\pm$  s.d.):  $72.08 \pm 5.48$ ,  $t(39)=25.15$ ; VLS: A1:  $61.11 \pm 2.15$ ,  $t(39)=32.25$ ; pTVA:  $63.89 \pm 2.78$ ,  $t(39)=31.22$ ), age (LIN: pTVA:  $54.58 \pm 4.14$ ,  $t(39)=6.90$ ; aTVA:  $63.96 \pm 12.55$ ,  $t(39)=6.94$ ; VLS: pTVA:  $65.00 \pm 7.26$ ,  $t(39)=12.89$ ; aTVA:  $60.42 \pm 5.19$ ,  $t(39)=12.54$ ) and identity (LIN: A1:  $9.20 \pm 9.23$ ,  $t(39)=2.24$ ; pTVA:  $9.48 \pm 4.90$ ,  $t(39)=4.59$ ; aTVA:  $9.41 \pm 6.28$ ,  $t(39)=3.51$ ; VLS: pTVA:  $16.18 \pm 7.05$ ,  $t(39)=9.11$ ; aTVA:  $8.23 \pm 4.70$ ,  $t(39)=3.12$ ) (**Fig. 5a-c**; Supplementary Tables 8-10).

Two-way ANOVAs with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on classification accuracy scores (gender, age, identity) revealed for gender classifications significant effects of Feature  $F(1, 312)=12.82$ ,  $p<.0005$  and ROI (gender:  $F(3, 312)=245.06$ ,  $p<.0001$ ; age:  $F(3, 312)=64.49$ ,  $p<.0001$ ; identity:  $F(3, 312)=14.49$ ,  $p<.0001$ ), as well as Feature x ROI interactions (gender:  $F(3, 312)=56.74$ ,  $p<.0001$ ; age:  $F(3, 312)=4.31$ ,  $p<.001$ ; identity:  $F(3, 312)=8.82$ ,  $p<.0001$ ). Post-hoc paired t-tests indicated that the VLS was better than LIN in preserving gender, age and identity information in at least one TVA compared with A1 (gender: aTVA:  $t(39)=5.13$ ,  $p<.0001$ ; age: pTVA:  $t(39)=9.78$ ,  $p<.0001$ ; identity: pTVA:  $t(39)=4.01$ ,  $p<.0005$ ) (all tests in Supplementary Table 11). Post-hoc two sample t-tests comparing ROIs revealed significant differences in all classifications, in particular with pTVA outperforming other ROIs in gender (LIN: pTVA vs A1:  $t(78)=22.40$ ,  $p<.0001$ ; pTVA vs mTVA:  $t(78)=10.92$ ,  $p<.0001$ ; pTVA vs aTVA:  $t(78)=31.47$ ,  $p<.0001$ ; VLS: pTVA vs A1:  $t(78)=4.94$ ,  $p<.0001$ ; pTVA vs mTVA:  $t(78)=13.96$ ,  $p<.0001$ ; pTVA vs aTVA:  $t(78)=22.06$ ,  $p<.0001$ ), age (LIN: pTVA vs A1:  $t(78)=7.26$ ,  $p<.0001$ ; pTVA vs mTVA:  $t(78)=10.11$ ,  $p<.0001$ ; VLS: pTVA vs A1:  $t(78)=5.71$ ,  $p<.0001$ ; pTVA vs mTVA:  $t(78)=10.11$ ,  $p<.0001$ ; pTVA vs aTVA:  $t(78)=3.21$ ,  $p<.005$ ) and identity (LIN: pTVA vs mTVA:  $t(78)=2.27$ ,  $p<.05$ ; VLS: pTVA vs A1:  $t(78)=6.45$ ,  $p<.0001$ ; pTVA vs mTVA:  $t(78)=6.62$ ,  $p<.0001$ ; pTVA vs aTVA:  $t(78)=5.85$ ,  $p<.0001$ ) (Supplementary Table 12).

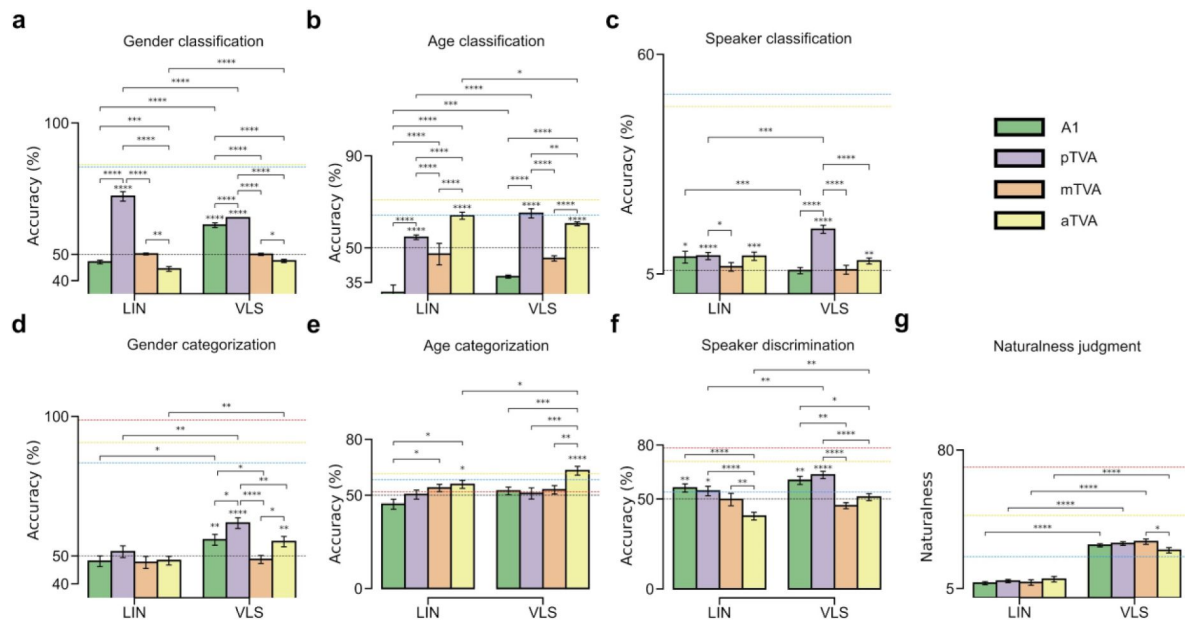
We further evaluated voice identity information in the reconstructed stimuli by testing human participants ( $n=13$ ) in a series of 4 online experiments assessing the reconstructed stimuli on (i) naturalness judgment, (ii) gender categorization, (iii) age categorization, and (iv) speaker categorization (cf. Methods). The naturalness rating task showed that the VLS-reconstructed stimuli sounded more natural compared to LIN-reconstructed ones, as revealed by a two-way repeated-measures ANOVA (factors: Feature and ROI) with a strong effect of Feature ( $F(1, 12)=53.72$ ,  $p<.0001$ ) and a small ROI x Feature interaction ( $F(3, 36)=5.36$ ,  $p<.005$ ). Post-hoc paired t-tests confirmed the greater naturalness of VLS-reconstructed stimuli in both A1 and the TVAs (all  $p<.0001$ ) (**Fig. 5g**). For the gender task, one-sample t-tests showed that categorization of the



**Fig. 4.**

### Reconstructing voice identity from brain recordings.

**a**, A linear voxel-based decoding model was used to predict the VLS coordinates of 18 Test Stimuli based on fMRI responses to ~12,000 Train stimuli in the different ROIs. To reconstruct the audio stimuli from the brain recordings, the predicted VLS coordinates were then fed to the trained decoder to yield reconstructed spectrograms, synthesized into sound waveforms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983 [\[4\]](#)). **b**, Reconstructed spectrograms of the stimuli presented to the participants. The left panels show the spectrogram of example original stimuli reconstructed from the VLS, and the right panels show brain-reconstructed spectrograms via LIN and the VLS (cf. Supplementary Audio 2).



**Fig. 5.**

### Behavioural and machine classification of the reconstructed stimuli.

**a,b,c,** Decoding voice identity information in brain-reconstructed spectrograms. Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (chance level: 50%), and identity (17 identities, chance level: 5.88%). Error bars indicate s.e.m across 40 random classifier initializations per ROI (instance of classifiers; 2 hemispheres x 20 seeds). The horizontal black dashed line indicates the chance level. The blue and yellow dashed lines indicate the LIN and VLS ceiling levels, respectively. \* $p < .05$ ; \*\* $p < .001$ ; \*\*\* $p < .001$ ; \*\*\*\* $p < .0001$ . **d,e,f,** Listener performance at categorizing speaker gender (chance level: 50%) and age (chance level: 50%), and at identity discrimination (2 forced choice task, chance level: 50%) in the brain-reconstructed stimuli. Error bars indicate s.e.m across participant scores. The horizontal black dashed line indicates the chance level, while the red, blue, and yellow dashed lines indicate the ceiling levels for the original stimuli, the LIN-reconstructed and the VLS-reconstructed, respectively. \* $p < .05$ ; \*\* $p < .01$ ; \*\*\* $p < .001$ ; \*\*\*\* $p < .0001$ . **g,** Perceptual ratings of voice naturalness in the brain-reconstructed stimuli' as assessed by human listeners, between 0 and 100 (zoomed between 5-80). \* $p < .05$ , \*\*\*\* $p < .0001$ .

reconstructed stimuli was only significantly above chance level for the VLS (A1: (mean accuracy  $\pm$  s.d.)  $55.77 \pm 10.84$ ,  $t(25)=2.66$ ,  $p<.01$ ; pTVA:  $61.75 \pm 7.11$ ,  $t(25)=8.26$ ,  $p<.0001$ ; aTVA:  $55.13 \pm 9.23$ ,  $t(25)=2.78$ ,  $p<.01$ ). Regarding the age and speaker categorizations, results also indicated that both the LIN- and VLS-reconstructed stimuli yielded above-chance performance in the TVAs (age: LIN: aTVA,  $55.77 \pm 14.95$ ,  $t(25)=1.93$ ,  $p<.05$ ; VLS: aTVA,  $63.14 \pm 11.82$ ,  $t(25)=5.56$ ,  $p<.0001$ ; identity: LIN: pTVA:  $54.38 \pm 9.34$ ,  $t(17)=1.93$ ,  $p<.05$ ; VLS: pTVA:  $63.33 \pm 6.75$ ,  $t(17)=8.14$ ,  $p<.0001$ ) (Supplementary Tables 13-15). Two-way repeated-measures ANOVAs revealed a significant effect of ROI for all categories (gender:  $F(3, 27)=5.90$ ,  $p<.05$ ; age:  $F(3, 36)=14.25$ ,  $p<.0001$ ; identity:  $F(3, 24)=38.85$ ,  $p<.0001$ ), and a Feature effect for gender ( $F(1, 9)=43.61$ ,  $p<.0001$ ) and identity ( $F(1, 8)=14.07$ ,  $p<.001$ ), but not for age ( $F(1, 12)=4.01$ ,  $p=0.07$ ), as well as a ROI x Feature interaction for identity discrimination ( $F(3, 24)=3.52$ ,  $p<.05$ ) (Supplementary Tables 16-17 for the model and ROI comparisons).

## Discussion

In this study we examined to what extent the cerebral activity elicited by brief voice stimuli can be explained by machine-learned representational spaces, specifically focusing on identity-related information. We trained a linear model and a DNN model to reconstruct 100,000s of short voice samples from 100+ speakers, providing low-dimensional spaces (LIN and VLS), which we related to fMRI measures of cerebral response to thousands of these stimuli. We find: (i) that 128 dimensions are sufficient to explain a sizeable portion of the brain activity elicited by the voice samples and yield brain-based voice reconstructions that preserve identity-related information; (ii) that the DNN-derived VLS outperforms the LIN space, particularly in yielding more brain-like representational spaces and more naturalistic voice reconstructions; (iii) that different ROIs have different degrees of brain-model relationship, with marked differences between A1 and the a, m, and pTVAs.

Low-dimensional spaces generated by machine learning have been used to approximate cerebral face representations and reconstruct recognizable faces based on fMRI (VanRullen & Reddy, 2019; Dado et al., 2022). In the auditory domain, however, they have mainly been used with a focus on linguistic (speech) information, ignoring identity-related information (but see Akbari et al., 2019). Here, we applied them to brief voice stimuli—with minimal linguistic content but already rich identity-related information—and found that as little as 128 dimensions account reasonably well for the complexity of cerebral responses to thousands of these voice samples as measured by fMRI (Fig. 2). LIN and VLS both showed brain-like representational geometries, particularly the VLS in the aTVAs (Fig. 3). They made possible what is, to our knowledge, the first fMRI-based voice reconstructions to preserve voice-related identity information such as gender, age, or even individual identity, as indicated by above-chance categorization or discrimination performance by both machine classifiers (Fig. 5a-c) and human listeners (Fig. 5d-f).

Estimation of fMRI responses (encoding) by LIN yielded correlations largely comparable to those by VLS (Fig. 2b), although many voxels were only explained by one or the other space (Fig. 2c). However, in the RSA, VLS yielded higher overall correlations with brain RDMs (Fig. 3), suggesting a representational geometry closer to that instantiated in the brain than LIN. Further, VLS-reconstructed stimuli sounded more natural than the LIN-reconstructed ones (Fig. 5g) and yielded both the best speaker discrimination by listeners (Fig. 5f) and speaker classification by machine classifiers (Fig. 5c). Unlike LIN, which was generated via linear transforms, VLS was obtained through a series of nonlinear transformations (Wetzel, 2017). The fact that the VLS outperforms LIN in decoding performance indicates that nonlinear transformation is required to better account for the brain representation of voices (Naselaris et al., 2011; Cowen et al., 2014; Han et al., 2019).



Comparisons between ROIs revealed important differences between A1 and the a, m, and pTVAs. For both LIN and VLS, fMRI signal (encoding) predictions were more accurate for the mTVAs than for A1, and for A1 than for the pTVAs (**Fig. 2b**). The aTVAs yielded the highest correlations with the models in the RSA (**Fig. 3**). Stimulus reconstructions (**Fig. 4**) based on the TVAs also yielded better gender, age, and identity classification than those based on A1, with gender and identity best preserved in the pTVA-, and to a lesser extent, in the aTVA-based reconstructions (**Fig. 5**). These results show that the a and pTVAs not only respond more strongly to vocal sounds than A1, but they also represent identity-related information in voice better than mTVA, which was previously anticipated in some neuroimaging studies (Gender: Charest et al., 2013; Identity: Belin & Zatorre, 2003; Maguinness et al., 2018; Roswandowitz et al., 2018; Aglieri et al., 2021). Moreover, several recent studies, using intracranial recordings, either through ECoG electrode grids (Zhang et al., 2021) or sEEG recordings (Rupp et al., 2022), found evidence that supports the idea of a hierarchical organization of voice patches in the temporal lobe, where the information flow starts from the mTVA patches and moves in two directions: one from mTVA to the anterior TVA (aTVA) and the other one from mTVA to posterior TVA (pTVA).

Overall, we show that a DNN-derived representational space provides an interesting approximation of the cerebral representations of brief voice stimuli that can preserve identity-related information. We find it remarkable that such results could be obtained to explain sound representations despite the poor temporal resolution of fMRI. Future work combining more complex architectures to time-resolved measures of cerebral activity, such as magnetoencephalography (Défossez et al., 2023) or ECoG (Pasley et al., 2012), will likely yield better models of the cerebral representations of voice information.

## Methods

### Experimental procedure overview

Three participants attended 13 MRI sessions each. The first session was dedicated to acquire high-resolution structural data, as well as to identify the voice-selective areas of each participant using a ‘voice localizer’ based on different stimuli than those in the same experiment (Pernet et al., 2015; see below).

Functional scanning was done using a rapid event-related design with a jittered inter-stimulus-interval (2.8-3.2 s). The next 12 sessions began with the acquisition of two fast structural scans for inter-session realignment purposes, followed by six functional runs, during which the main stimulus set of the experiment was presented. Each functional run lasted approximately 12 minutes. Participants 1 and 2 attended all scanning sessions (72 functional runs in total); due to technical issues, Participant 3 only performed 24 runs.

Participants were instructed to stay in the scanner while listening to the stimuli. To maintain participants’ awareness during functional scanning, they were asked to press an MRI-compatible button each time they heard the same stimulus two times in a row, a rare event occurring 3% of the time (correct button hits (median accuracy  $\pm$  s.d.): S1=96.67 $\pm$ 7.10, S2=100.00 $\pm$ 0.89, S3=95.00 $\pm$ 3.68).

Scanning sessions were spaced by at least two days to avoid possible auditory fatigue due to the exposure to scanner noise. To ensure that participants’ hearing abilities did not vary across scanning sessions, hearing thresholds were measured before each session using a standard audiometric procedure (Martin & Champlin, 2000; ISO 2004) and compared with the thresholds obtained prior the first session.

## Participants

This study was part of the project ‘Réseaux du Langage’ and was promoted by the National Center for Scientific Research (CNRS). It has been given favorable approval by the local ethics committee (Comité de Protection des Personnes Sud-Méditerranée) on the date of 13th February 2019. The National Agency for Medicines (ANSM) has been informed of this study, which is registered under the number 2017-A03614-49. Three native French human speakers were scanned (all females; 26–33 years old). Participants gave written informed consent and received a compensation of 40€ per hour for their participation. All were right-handed and no one had hearing disorder or neurological disease. All participants had normal hearing thresholds of 15 dB HL, for octave frequencies between 0.125 and 8 kHz.

## Stimuli

The auditory stimuli were divided into two sequences. One ‘voice localizer’ sequence to identify the voice-selective areas of each participant (Pernet et al., 2015 [DOI](#)) and a main voice stimuli.

### Voice localizer stimuli

The voice localizer stimuli consisted of 96 complex sounds of 500ms grouped in four categories of human voice, macaque vocalizations, marmoset vocalizations, and complex non-vocal sounds (more details in Bodin et al., 2021 [DOI](#)).

### Main voice stimuli

The main stimulus set consisted of brief human voice sounds sampled from the Common Voice dataset (Ardila et al., 2020 [DOI](#)). Stimuli were organized into four main category levels: language (English, French, Spanish, Deutch, Polish, Portuguese, Russian, Chinese), gender (female/male), age (young/adult; young: teenagers and twenties; adult: thirties to sixties included) and identity (S1: 135 identities; S2: 142 identities; S3: 128 identities; ~44 samples per identity). Throughout the manuscript, the term ‘gender’ rather than ‘sex’ was utilized in reference to the demographic information obtained from the participants of the Common Voice dataset (Ardila et al., 2020 [DOI](#)), as it was the terminology employed in the survey (‘male/female/other’). Stimulus sets were different for each participant and the number of stimuli per set also varied slightly (number of unique stimuli: Participant 1, N=6150; Participant 2, N=6148; Participant 3, N=5123). For each participant, six stimuli were selected randomly among the sounds having a high energy (as measured with the amplitude envelope) from their stimulus set and were repeated extensively (60 times), to improve the performance of the brain decoding (VanRullen & Reddy, 2019 [DOI](#); Horikawa & Kamitani, 2017 [DOI](#); Chang et al., 2019 [DOI](#)); these will be called the “repeated” stimuli hereafter, the remaining stimuli were presented twice. The third participant attended 5 BrainVoice sessions instead of 12, one BrainVoice session corresponding to 1030 stimuli (1024 unique stimuli and 6 ‘test’ stimuli). Specifically, 5270 stimuli were presented to the third participant instead of ~12,000 for the two others. Among these 5270 stimuli, 5120 unique stimuli were presented once, as for the two other participants, 6 ‘test’ stimuli were presented 25 times (150 trials). The stimuli were balanced within each run according to language, gender, age, and identity, as to avoid any potential adaptation effect. In addition, identity was balanced across sessions.

All stimuli of the main set were resampled at 24414 Hz and adjusted in duration (250 ms). For each stimulus, a fade-in and a fade-out were applied with a 15 ms cosine ramp to their onset and offset, and were normalized by dividing the root mean square amplitude. During fMRI sessions, stimulus presentations were controlled using custom Matlab scripts (Mathworks, Natick, MA, USA) interfaced with an RM1 Mobile Processor (Tucker-David Technologies, Alachua, USA). The auditory stimuli were delivered pseudo-randomly through MRI-compatible earphones (S14, SensiMetrics, USA) at a comfortable sound pressure level that allowed for clear and intelligible listening.

## Computational models

We used two computational models to learn representational space for voice signals, Linear Autoencoder (LIN) and Deep Variational Autoencoder (VAE; Kingma & Welling, 2014). Both are encoder-decoder models that are learnt to reproduce at their output their input while going through a low dimensional representation space usually called latent space (that we will call *voice latent space* since they are learnt on voice data). The autoencoders were trained on a dataset of 182K sounds from the Common Voice dataset (Ardila et al., 2020), balanced in gender, language and identity to reduce the bias in the synthesis (Gutierrez et al., 2021). Both models operate on sounds which were represented as spectrograms that we describe below. These representations were tested in all the encoding/decoding and RSA analyses.

## Spectrograms

We used amplitude spectrograms as input of the models that we describe below. Short term Fourier transforms of the waveform were computed using a sliding window of length 50 ms with a hop size of 12.5 ms (hence an overlap of 37.5 ms) and applying a Hamming window of size 800 samples before computing the Fourier transform of each slice. Only the magnitude of the spectrogram was kept and the phase of the complex representation was removed. At the end, a 250 ms sound is represented by a  $21 \times 401$  matrix with 21 time steps and 401 frequency bins.

We used a custom code based on *numpy.fft* package (Harris et al., 2020). The size and the overlap between the sliding windows of the spectrogram were chosen to conform with the uncertainty principle between time and frequency resolution. The main constraint was to find a trade-off between accurate phase reconstruction with the Griffin & Lim algorithm (1983) and a reasonable size of the spectrogram.

We standardized each of the 401 frequency bands separately, by centering all the data corresponding to each frequency band at every time step in all spectrograms, which involved removing their mean, and dividing by their standard deviation. This separate standardization of frequency bands resulted in a smaller reconstruction error compared to standardizing across all the bands.

## Deep neural network

We designed a deep variational autoencoder (VAE; Kingma & Welling, 2014) of 15 layers with an intermediate hidden representation of 128 neurons that we refer to as the *voice latent space* (VLS). In an autoencoder model, the two sub-network components, the *Encoder* and the *Decoder*, are jointly learned on complementary tasks (Fig. 1a). The Encoder network (noted *Enc* hereafter; 7 layers) learns to map an input,  $s$  (a spectrogram of a sound), onto a (128-dimensional) *voice latent space* representation ( $z$ ; in blue in the middle of Fig. 1a), while the Decoder (noted *Dec* hereafter; 7 layers) aims at reconstructing the spectrogram  $s$  from  $z$ . The learning objective of the full model is to make the output spectrogram  $Dec(Enc(s))$  as close as possible to the original one  $s$ . This reconstruction objective is defined as the L2 loss,  $\|Dec(Enc(s)) - s\|^2$ . The parameters of the Encoder and of the Decoder are jointly learned using gradient descent to optimize the average L2 loss computed on the training set  $\sum_{s \in \text{Training Set}} \|Dec(Enc(s)) - s\|^2$ . We trained this DNN on the Common Voice dataset (Ardila et al., 2020) according to VAE learning procedure (as explained in Kingma & Welling, 2019) until convergence (network architecture and particularities of the training procedure are provided in Supplementary Table 1), using the PyTorch python package (Paszke et al., 2019).

## Linear autoencoder

We trained a linear autoencoder on the same dataset (described above) to serve as a linear baseline. Both the *Encoder* and the *Decoder* networks consisted of a single fully-connected layer, without any activation functions. Similar to the VAE, the latent space obtained from the *Encoder* was a 128-dimensional vector. The parameters of both the *Encoder* and of the *Decoder* were jointly learned using gradient descent to optimize the average L2 loss computed on the training set.

## Neuroimaging data acquisition

Participants were scanned using a 3 Tesla Prisma scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel receiver head-coil. Their movements were monitored during the acquisition using the software FIRMM (Dosenbach et al., 2017). The whole-head high-resolution structural scan acquired during the first session was a T1-weighted multi-echo MPAGE (MEMPAGE) (TR = 2.5 s, TE = 2.53, 4.28, 6.07, 7.86 ms, TI=1000 ms flip angle: 8°, matrix size = 208 × 300 × 320; resolution 0.8 × 0.8 × 0.8 mm<sup>3</sup>, acquisition time: 8min22s). Lower resolution scans acquired during all other sessions were T1-weighted MPAGE scans (TR = 2.3 s, TE = 2.88 ms, TI=900ms, flip angle: 9°, matrix size = 192 × 240 × 256; resolution 1 × 1 × 1 mm<sup>3</sup>, sparse sampling with 2.8 times undersampling and compressed sensing reconstruction, acquisition time: 2min37). Functional imaging was performed using an EPI sequence (multiband factor = 5, TR = 462 ms, TE = 31.2 ms, flip angle: 45°, matrix size = 84 × 84 × 35, resolution 2.5 × 2.5 × 2.5 mm<sup>3</sup>). Functional slices were oriented parallel to the lateral sulci with a z-axis coverage of 87.5 mm, allowing it to fully cover both the TVAs (Pernet et al., 2015) and the FVAs (Aglieri et al., 2018). The physiological signals (heart rate and respiration) were measured with the external sensors of Siemens.

## Pre-processing of neuroimaging data and general linear modeling

Tissue segmentation and brain extraction was performed on the structural scans using the default segmentation procedure of SPM 12 (Ashburner et al., 2012). The preprocessing of the BOLD responses involved correcting motion, registering inter-runs, detrending and smoothing the data. Each functional volume was realigned to a reference volume taken from a steady period in the session that was spatially the closest to the average of all sessions.

Transformation matrices between anatomical and functional data were computed using boundary-based registration (FSL; Smith et al., 2004). The data were respectively detrended and smoothed using the *nilearn* functions *clean\_img* and *smooth\_img* (kernel size of 3mm) (Abraham et al., 2014), resulting in the matrix  $Y \in R^{S \times V}$ , with  $S$  the number of scans and  $V$  the number of voxels.

A first general linear model (GLM) was fit to regress out the noise by predicting  $Y$  from a “denoised” design matrix, composed of  $R = 38$  regressors of nuisance (Supplementary Figure 3). These regressors of nuisance, also called covariates of no interest, included: 6 head motion parameters (3 variable for the translations, 3 variables for the rotations); 18 ‘RETROICOR’ regressors (Glover et al., 2000) using the *TAPAS PhysIO* package (Kasper et al., 2017) (with the hyperparameters set as specified in Snoek et al.) were computed from the physiological signals; 13 regressors modeling slow artifactual trends (sines and cosines, cut frequency of the high-pass filter = 0.01 Hz); and a confound-mean predictor. The design matrix was convolved with an hemodynamic response function (HRF) with a peak at 6 s and an undershoot at 16 s (Glover et al., 1999), we note the convolved design matrix as  $X_d \in R^{S \times R}$ . The “denoise” GLM’s parameters  $\beta_d \in R^{R \times V}$  were optimized to minimize the amplitude of the residual  $\beta_d = \operatorname{argmin}_{\beta \in R^{R \times V}} ||Y - X_d \beta||^2$ . We used a lag-1 autoregressive model (ar(1)) to model the temporal structure of the noise (Friston et al., 2002). The *denoised* BOLD signal  $Y_d$  was then obtained from the original one according to  $Y_d = Y - (X_d \beta_d) \in R^{S \times V}$ .



A second “stimulus” GLM model was used to predict the denoised BOLD responses for each stimulus using a design matrix  $X_s \in R^{S \times (N_s+1)}$  (which was convolved with an hemodynamic response function, HRF as above) and a parameters matrix  $\beta_s \in R^{(N_s+1) \times V}$  where  $N_s$  stands for the number of stimuli. The last row (resp. column) of  $\beta_s$  (resp.  $X_s$ ) stands for a silence condition. Again,  $\beta_s$  was learned to minimize the residual  $\beta_s = \operatorname{argmin}_{\beta \in R^{(N_s+1) \times V}} ||Y_d - X_s \beta||^2$ . Once learned, each of the first  $N_s$  line of  $\beta_s$  was corrected by subtracting the  $(N_s+1)^{th}$  line, yielding the contrast maps for stimuli  $\tilde{\beta}_s \in R^{N_s \times V}$ . We note hereafter  $\tilde{\beta}_s[i,:]$   $\in R^V$  the contrast map for a given stimulus, it is the  $i^{th}$  line of  $\tilde{\beta}_s$ .

A third “identity” GLM was fit to predict the BOLD responses of each voice speaker identity, using a design matrix  $\beta_i \in R^{(N_i+1) \times V}$  and a design matrix  $X_i \in R^{S \times (N_i+1)}$  (which was again convolved with an hemodynamic response function, HRF) where  $N_i$  stands for the number of unique speakers. Again the last row/column in  $\beta_i$  and  $X_i$  stands for the silent condition.  $\beta_i$  is learned to minimize the residual  $\beta_i = \operatorname{argmin}_{\beta \in R^{(N_i+1) \times V}} ||Y_d - X_i \beta||^2$  (Supplementary Figure 2a). Again, the final speaker contrast maps were obtained by contrasting (i.e., subtracting) the regression coefficients in a row of  $\beta_i$  with the silence condition (last row; Supplementary Figure 2a), yielding  $\tilde{\beta}_i \in R^{N_i \times V}$ . Here the  $j^{th}$  row of  $\tilde{\beta}_i$ ,  $\tilde{\beta}_i[j,:]$   $\in R^V$ , represents the amplitude of the BOLD response of the contrast map for speaker  $j$  (i.e. to all the stimuli from this speaker).

A fourth “localizer” GLM model was used to predict the denoised BOLD responses of each sound category from the *Voice localizer stimuli* presented above. The procedure was similar as described for the two previous GLM models. Once the GLM was learned, we contrasted the human voice category with the other sound categories in order to localize for each participant the posterior Temporal Voice Area (pTVA), medial Temporal Voice Area (mTVA) and anterior Temporal Voice Area (aTVA) in each hemisphere. The center of each TVA corresponded to the local maximum of the voice > non voice t-map whose coordinates were the closest to the TVAs reported in (Pernet et al., 2015 [\[15\]](#)). The analyses were carried on for each region of interest (ROI) of each hemisphere.

Additionally, we defined for each participant the primary auditory cortex (A1) as the maximum value of the probabilistic map (non-linearly registered to each participant functional space) of Heschl's gyri provided with the MNI152 template (Penhune et al., 1996 [\[16\]](#)), intersected with the sound vs silence contrast map.

## Identity-based and stimulus-based representations

We performed analyses either at the stimulus level, e.g. predicting the neural activity of a participant listening to a given *stimulus* ( $\tilde{\beta}_s$ 's lines) from the *voice latent space* representation of this stimuli, or at the speaker identity level, e.g. predicting the average neural activity in response to stimuli of a given speaker identity ( $\tilde{\beta}_i$ 's lines) from this speaker's *voice latent space* representation. The identity-based analyses were used for the characterization of the *voice latent space* (Fig. 1 [\[17\]](#)), the brain encoding (Fig. 2 [\[18\]](#)), and the representational similarity analysis (Fig. 3 [\[19\]](#)), while the stimulus-based analyses were used for the brain decoding analyses (Fig. 4 [\[20\]](#), 5).

We conducted stimulus-based analyses to examine the relationship between stimulus contrast maps in neural activity ( $\tilde{\beta}_s$ ) and the encodings of individual stimulus spectrograms computed by the encoder of an autoencoder model (either linear or deep variational autoencoder) on the computational side. We will note  $z_s^{lin} \in R^{N_s \times 128}$  encodings of stimuli by the LIN model and  $z_s^{vae} \in R^{N_s \times 128}$  the encodings of stimuli computed by the VAE model. The encoding of the  $k^{th}$  stimuli by one of these models is the  $k^{th}$  row of the corresponding matrix and it is noted as  $z_s^{model}[k,:]$ .

For identity-based analyses we studied relationships between identity contrast maps in  $\tilde{\beta}_i$  on the neural activity side, and an encoding of speaker identity in the VLS implemented by an autoencoder model (LIN or VAE) on the computational side, e.g. we note  $z_i^{vae}[j]$  the representation of speaker  $j$  as computed by the *vae* model. We chose to define a speaker identity-based representation as the average of a set of sample-based representations for stimuli from this

speaker, e.g.  $z_i^{model}[j] = 1/|S_j| \sum_{k \in S_j} z_k^{model}[i,:]$  where  $S_j$  stands for the set of stimuli by speaker  $j$  and  $model$  stands for *vae* or *lin*. Averaging in the *voice latent space* is expected to be much more powerful and relevant than averaging in the input space spectrograms (VanRullen & Reddy, 2019 [DOI](#)).

## Characterization of the autoencoder latent space

We characterized the organization of the *voice latent space* (VLS) and of the features computed by the linear autoencoder (LIN) by measuring through classification experiments the presence of information about speaker's gender, age, and identity in the representations learned by these models.

We first computed the speaker's identity *voice latent space* representations for each of the 405 speakers in the main voice dataset (135+142+128 see *Stimuli* section) as explained above.

Next we used these speakers' *voice latent space* representation to investigate if the gender, age, identity were encoded in the VLS. To do so we divided the data in separate train and test sets and learned classifiers to predict gender, age, or identity from the train set. The balanced (to avoid the small effects associated with unbalanced folds) accuracy of the classifiers were then evaluated on the test set. The higher the performance on the test set the more we are confident that the information is encoded in the VLS. More specifically for each task (gender, age, identity), we trained a Logistic Regression classifier (linear regularized logistic regression; L2 penalty,  $\text{tol}=0.0001$ ,  $\text{fit\_intercept}=\text{True}$ ,  $\text{intercept\_scaling}=1$ ,  $\text{max\_iter}=100$ ) using the scikit-learn python package (Pedregosa et al., 2018 [DOI](#)).

In order to statistically evaluate the significance of the results and to avoid a potential overfitting, the classifications were repeated 20 times with 20 different initializations (*seed*) and the metrics were then averaged for each voice category (gender, age). More specifically, we repeated the following experiment 20 times with 20 different random seeds. For each seed, we performed 5 train-test splits with 80% of the data in the training and 20% in the test set. For each split we used 5-fold cross validation on the training set to select the optimal value for the regularization hyperparameter  $C$  (searching between 10 values logarithmically spaced on the interval  $[-3, +3]$ ). We then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results were then averaged over 20 experiments. Note that data were systematically normalized with a scaler fitted on the training set. We used a robust scaling strategy for these experiments (removing the median, then scaling to the quantile range; 25<sup>th</sup> quantile and 75<sup>th</sup> quantile) which occurs to be more relevant with a small training set.

To investigate how speaker identity information is encoded in the latent space representations of speakers' voices, we computed speaker identity *voice latent space* representations by averaging 20 stimulus-based representations, in order to obtain a limited amount of data per identity that could be distributed across training and test datasets.

We first tested whether the mean of the distribution of accuracy scores obtained for 20 seeds was significantly above chance level using one-sample t-tests. We then evaluated the difference in classification accuracy between the VLS and LIN via one-way ANOVAs (dependent variable: test balance accuracy; between factor: Feature), for each category (speaker gender, age, identity). We performed post-hoc planned paired t-tests between the models to test the significance of the VLS-LIN difference.

## Brain encoding

We performed encoding experiments on identity-based representations for each of the three participants (Fig. 2). For each participant we explored the ability to learn a regularized linear regression that predicts a speaker-based neural activity, e.g. the  $j^{\text{th}}$  speaker's contrast map  $\tilde{\beta}_i[j] \in R^V$ , from this speaker's voice latent space representation, that we note  $z_i^{\text{model}}[j] \in R^{128}$  (Fig. 2a). We carried out these regression analyses for each ROI (A1, pTVA, mTVA, aTVA) in each hemisphere and participant, independently.

The regression model parameters  $\hat{W}_{\text{encod}} \in R^{128 \times V}$  were learned according to:

$$\hat{W}_{\text{encod}} = \underset{W_{\text{encod}} \in R^{128 \times V}}{\text{argmin}} \sum_{j=1, N_i} (z_i^{\text{model}}[j] \times W_{\text{encod}} - \tilde{\beta}_i[j])^2 + \lambda \|W_{\text{encod}}\|^2$$

where  $\lambda$  is a hyperparameter tuning the optimal tradeoff between the data fit and the penalization terms above. We used the ridge regression with built-in cross-validation as implemented as *RidgeCV* in the scikit-learn library (Pedregosa et al., 2018).

The statistical significance of each result was assessed with the following procedure. We repeated the following experiment 20 times with different random seeds. Each time, we performed 5 train-test splits with 80% of the data in the training and 20% in the test set. For each split we used *RidgeCV* (relying on leave-one-out) on the training set to select the optimal value for the hyperparameter  $\lambda$  (searching between 10 values logarithmically spaced on the interval  $[10^{-1}; 10^8]$ ). Following standard practice in machine learning, we then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results are then averaged over 20 experiments. Note that here again with small training sets data were systematically normalized in each experiment using robust scaling.

Evaluation relied on the 'brain score'-inspired procedure (Schrimpf et al., 2018) which evaluates the performance of the ridge regression with a Pearson's correlation score. Correlations between measured neural activities  $\tilde{\beta}_i$  and predicted ones  $z_i^{\text{model}} * \hat{W}_{\text{encod}}$  were computed for each voxel and averaged over repeated experiments (folds and seeds) yielding one correlation value for every voxel and for every setting. The significance of the results was assessed with one-sample t-tests for the Fisher z-transformed correlation scores (3 x participants x 2 hemispheres x V voxels). For each region of interest, the scores are reported across participants and hemispheres (Fig. 2b). The exact same procedure was followed for the LIN modeling.

In order to determine which of the two feature spaces (VLS, LIN) and which of the two ROI (A1, TVAs) yielded the best prediction of neural activity, we compared the means of distributions of correlations coefficients using a mixed ANOVA performed on the Fisher z-transformed coefficients (dependent variable: correlation; between factor: ROI; repeated measurements: Feature; between-participant identifier: voxel).

For each ROI, we then used t-tests to perform post-hoc contrasts for the VLS-LIN difference in brain encoding performance (comparison tests in Fig. 2b; Supplementary Table 4). We finally conducted two-sample t-tests between the brain encoding model's scores trained to predict A1 and those trained to predict temporal voice areas to test the significance of the A1-TVAs difference (Supplementary Table 5).

The statistical tests were all performed using the *pingouin* python package (Vallat, 2018).

## Representational similarity analysis

The RSA analyses were carried out using the package *rsatoolbox* (Schütt et al., 2021 <https://github.com/rsagroup/rsatoolbox>). For each participant, region of interest and hemisphere, we computed the cerebral Representational Dissimilarity Matrix (RDM) using the Pearson's correlation between the speaker identity-specific response patterns of the GLM estimates  $\tilde{\beta}_i$  (Walther et al., 2016 [\(Fig. 3a\)](#)). The model RDMs were built using cosine distance (Xing et al., 2015 [\(Fig. 3a\)](#); Bhattacharya et al., 2017 [\(Fig. 3a\)](#); Wang et al., 2018 [\(Fig. 3a\)](#)), capturing speaker pairwise feature differences predicted by the computational models LIN and the VLS (**Fig. 3a**). For greater comparability with the rest of the analyses described here, the GLM estimates and the computational models' features were first normalized using robust scaling. We computed the Spearman correlations coefficients between the brain RDMs for each ROI, and the two model's RDMs (**Fig. 3b**). We assessed the significance of these brain-model correlation coefficients within a permutation-based 'maximum statistics' framework for multiple comparison correction (one-tailed inference; N permutations = 10,000 for each test; permutation of rows and columns of distance matrices, see Giordano et al., 2023 [\(Fig. 3b\)](#) and Maris & Oostenveld, 2007; see **Fig. 3b**). We evaluated the VLS-LIN difference using a two-way repeated-measures ANOVA on the Fisher z-transformed Spearman correlation coefficients (dependent variable: correlation; within factors: ROI and Feature; participant identifier: participant hemisphere pair). The same permutation framework was also used to assess the significance of the difference between the RSA correlation for the VLS and LIN models.

## Brain decoding

Brain decoding was investigated at the stimulus level. The stimuli's voice latent space representations  $z_s^{model} \in R^{N \times 128}$  and voice samples' contrast maps  $\tilde{\beta}_s \in R^{N \times V}$  were divided into train and test splits, normalized across voice samples using robust scaling, then fit to the training set. For every participant and each ROI, we trained a  $L_2$ -regularized linear model  $W \in R^{V \times 128}$  model to predict the voice samples' latent vectors from the voice samples' contrast maps (**Fig. 4a**). The hyperparameter selection and optimization was done similarly as in the Brain encoding scheme. Training was performed on non repeated stimuli (see Stimuli section). We then used the trained models to predict for each participant the 6 repeated stimuli that were the most presented. Waveforms were estimated starting from the reconstructed spectrograms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983 [\(Fig. 4a\)](#)).

We then used classifier analyses to assess the presence of voice information (gender, age, speaker identity) in the reconstructed latent representations (i.e., the latent representation predicted from the brain activity of a participant listening to a specific stimulus) (**Fig. 5a, b, c**). To this purpose, we first trained linear classifiers to categorize the training voice stimuli (participant 1, N = 6144; participant 2, N = 6142; participant 3, N = 5117; total, N = 17403) by gender (2 classes), age (2 classes) or identity (17 classes) based on VLS coordinates.

Secondly, we used the previously trained classifiers to predict the identity information based on the VLS derived from the brain responses of the 18 Test voice stimuli (3 participants x 6 stimuli). We first tested using one-sample t-tests that the mean of the distribution of accuracy scores obtained across random classifier initializations of classifiers (2 hemispheres x 20 seeds = 40) was significantly above chance level, for each category, ROI and model. We then evaluated the difference in performance at preserving identity-related information depending on the model or ROI via two-way ANOVAs (dependent variable: accuracy; between factors: Feature and ROI). We performed post-hoc planned paired t-tests between each model pair to test the significance of the VLS-LIN difference. Two-sample t-tests were finally used to test the significance of the A1-TVAs difference.



## Listening tests

We recruited 13 participants through the online platform Prolific ([www.prolific.co](http://www.prolific.co)) for a series of online behavioral experiments. All participants reported having normal hearing. The purpose of these experiments was to evaluate how well voice identity information and naturalness are preserved in fMRI-based reconstructed voice excerpts. In the main session, participants carried out 4 tasks, in the following order: ‘speaker discrimination’ (~120 min), ‘perceived naturalness’ (~30 min), ‘gender categorization’ (~30 min), ‘age categorization’ (~30 min). The experiment lasted 3 hours and 35 minutes, and each participant was paid £48. 12 participants performed the speaker discrimination task, and all participants performed the other tasks.

Prior to the main experiment session, participants carried out a short loudness-change detection task to ensure that they wore headphones, and that they were attentive and properly set up for the main experiment (Woods et al., 2017). On each of 12 trials, participants heard 3 tones and were asked to identify which tone was the least loud by clicking on one of 3 response buttons: ‘First’, ‘Second’, or ‘Third’. Participants were admitted to the main experiment only if they achieved perfect performance in this task. We additionally refined the participant pool by excluding those who performed badly on the original stimuli, by retaining only the subjects whose performance was above the 25th percentile of accuracy. (gender and age categorizations: as all participants performed well (**Fig. 5d, e, red** dotted lines); speaker discrimination: 9/12 participants performed above the threshold of 64%).

The next three tasks were each carried out on the same set of 342 experimental stimuli, each presented on a different trial: 18 original stimuli, 36 stimuli reconstructed directly from the LIN and the VLS models, and 18 stimuli x 2 models x 4 regions of interest x 2 hemispheres= 288 brain-reconstructed stimuli.

In the ‘perceived naturalness’ task, participants were asked to rate how natural the voice sounded on a scale ranging from ‘Not at all natural’ to ‘Highly natural’ (i.e., similar to a real recording), and were instructed to use the full range of the scale.

During the ‘gender categorization’ task, participants categorized the gender by clicking on a ‘Female’ or ‘Male’ button.

Finally, in the ‘age categorization’ task, participants categorized the age of the speaker by clicking on a ‘Younger’ or ‘Older’ button.

In the ‘speaker discrimination’ task, participants carried out 684 trials (342 experimental stimuli x 2) with short breaks in between. On each trial, they were presented with 2 short sound stimuli, one after the other, and participants had to indicate whether they were from the same speaker or not. The speech material was selected randomly and was different between two stimuli.

To evaluate the performance of the participants, we firstly conducted one-sample t-tests to examine whether the mean accuracy score calculated from their responses was significantly higher than the chance level for each model and ROI. Next, we used two-way repeated-measures ANOVAs to assess the variation in participants’ performances in identifying identity-related information (dependent variable: accuracy; between-participant factors: Feature and ROI). To determine the statistical significance of the VLS-LIN difference, we carried out post-hoc planned paired t-tests between each model pair. Finally, we employed two-sample t-tests to evaluate the statistical significance of the A1-TVAs difference.

## Data and code availability

All data and codes will be made publicly available upon the article publication.

## Acknowledgements

We thank Bruno Nazarian for the design of an MRI-compatible button. We thank Jean-Luc Anton and Kepkee Loh for useful discussions. This work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 788240). This work was performed in the Center IRM-INT@CERIMED (UMR 7289, AMU-CNRS), platform member of France Life Imaging network (grant ANR-11-INBS-0 0 06). This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (*France 2030*), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A\*MIDEX).

## References

- Abraham Alexandre, Pedregosa Fabian, Eickenberg Michael, Gervais Philippe, Mueller Andreas, Kossaifi Jean, Gramfort Alexandre, Thirion Bertrand, Varoquaux Gael (2014) **Machine Learning for Neuroimaging with Scikit-Learn** *Frontiers in Neuroinformatics* **8**
- Aglieri Virginia, Cagna Bastien, Velly Lionel, Takerkart Sylvain, Belin Pascal (2021) **FMRI-Based Identity Classification Accuracy in Left Temporal and Frontal Regions Predicts Speaker Recognition Performance** *Scientific Reports* **11** <https://doi.org/10.1038/s41598-020-79922-7>
- Akbari Hassan, Khalighinejad Bahar, Herrero Jose L., Mehta Ashesh D., Mesgarani Nima (2019) **Towards Reconstructing Intelligible Speech from the Human Auditory Cortex** *Scientific Reports* **9** <https://doi.org/10.1038/s41598-018-37359-z>
- Ardila Rosana, Branson Megan, Davis Kelly, Henretty Michael, Kohler Michael, Meyer Josh, Morais Reuben, Saunders Lindsay, Tyers Francis M., Weber Gregor (2020) **Common Voice: A Massively-Multilingual Speech Corpus** *arXiv*
- Ashburner John (2012) **SPM: A History** *NeuroImage* **62**:791–800 <https://doi.org/10.1016/j.neuroimage.2011.10.025>
- Barbero Francesca M., Calce Roberta P., Talwar Siddharth, Rossion Bruno, Collignon Olivier, ENEURO.0471-20.2021 (2021) **Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain** *eNeuro* **8** <https://doi.org/10.1523/ENEURO.0471-20.2021>
- Belin Pascal, Bestelmeyer Patricia E. G., Latinus Marianne, Watson Rebecca (2011) **Understanding Voice Perception: Understanding Voice Perception** *British Journal of Psychology* **102**:711–25 <https://doi.org/10.1111/j.2044-8295.2011.02041.x>
- Belin Pascal, Bodin Clémentine, Aglieri Virginia (2018) **A ‘Voice Patch’ System in the Primate Brain for Processing Vocal Information?** *Hearing Research* **366**:65–74 <https://doi.org/10.1016/j.heares.2018.04.010>
- Belin Pascal, Fecteau Shirley, Bédard Catherine (2004) **Thinking the Voice: Neural Correlates of Voice Perception** *Trends in Cognitive Sciences* **8**:129–35 <https://doi.org/10.1016/j.tics.2004.01.008>
- Belin Pascal, Zatorre Robert J. (2003) **Adaptation to Speaker’s Voice in Right Anterior Temporal Lobe** *NeuroReport* **14**:2105–9 <https://doi.org/10.1097/00001756-200311140-00019>
- Belin Pascal, Zatorre Robert J., Lafaille Philippe, Ahad Pierre, Pike Bruce (2000) **Voice-Selective Areas in Human Auditory Cortex** *Nature* **403**:309–12 <https://doi.org/10.1038/35002078>
- Bhattacharya Gautam, Alam Jahangir, Kenny Patrick (2017) **Deep Speaker Embeddings for Short-Duration Speaker Verification** *Interspeech* :1517–21
- Blank Helen, Wieland Nuri, von Kriegstein Katharina (2014) **Person Recognition and the Brain: Merging Evidence from Patients and Healthy Individuals** *Neuroscience & Biobehavioral Reviews* **47**:717–34 <https://doi.org/10.1016/j.neubiorev.2014.10.022>

Bodin Clémentine, Trapeau Régis, Nazarian Bruno, Sein Julien, Degiovanni Xavier, Baurberg Joël, Rapha Emilie, Renaud Luc, Giordano Bruno L., Belin Pascal (2021) **Functionally Homologous Representation of Vocalizations in the Auditory Cortex of Humans and Macaques** *Current Biology* **31** <https://doi.org/10.1016/j.cub.2021.08.043>

Bouthillier Xavier *et al.* (2021) **Accounting for Variance in Machine Learning Benchmarks** *Proceedings of Machine Learning and Systems 3 (MLSys 2021)*

Capilla A., Belin P., Gross J. (2013) **The Early Spatio-Temporal Correlates and Task Independence of Cerebral Voice Processing Studied with MEG** *Cerebral Cortex* **23**:1388–95 <https://doi.org/10.1093/cercor/bhs119>

Caucheteux Charlotte, Gramfort Alexandre, King Jean-Rémi (2022) **Deep Language Algorithms Predict Semantic Comprehension from Brain Activity** *Scientific Reports* **12** <https://doi.org/10.1038/s41598-022-20460-9>

Caucheteux Charlotte, Gramfort Alexandre, King Jean-Rémi (2023) **Evidence of a Predictive Coding Hierarchy in the Human Brain Listening to Speech** *Nature Human Behaviour* :1–12 <https://doi.org/10.1038/s41562-022-01516-2>

Caucheteux Charlotte, King Jean-Rémi (2022) **Brains and Algorithms Partially Converge in Natural Language Processing** *Communications Biology* **5**:1–10 <https://doi.org/10.1038/s42003-022-03036-1>

Chang Nadine, Pyles John A., Marcus Austin, Gupta Abhinav, Tarr Michael J., Aminoff Elissa M. (2019) **BOLD5000, a Public fMRI Dataset While Viewing 5000 Visual Images** *Scientific Data* **6** <https://doi.org/10.1038/s41597-019-0052-3>

Charest I., Pernet C., Latinus M., Crabbe F., Belin P. (2013) **Cerebral Processing of Voice Gender Studied Using a Continuous Carryover fMRI Design** *Cerebral Cortex* **23**:958–66 <https://doi.org/10.1093/cercor/bhs090>

Charest Ian, Pernet Cyril R., Rousselet Guillaume A., Quiñones Ileana, Latinus Marianne, Fillion-Bilodeau Sarah, Chartrand Jean-Pierre, Belin Pascal (2009) **Electrophysiological Evidence for an Early Processing of Human Voices** *BMC Neuroscience* **10** <https://doi.org/10.1186/1471-2202-10-127>

Chen Zijiao, Qing Jiaxin, Xiang Tiange, Yue Wan Lin, Zhou Juan Helen, in 2023 (2023) **Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding** *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE :22710–20

Cowen Alan S., Chun Marvin M., Kuhl Brice A. (2014) **Neural Portraits of Perception: Reconstructing Face Images from Evoked Brain Activity** *NeuroImage* **94**:12–22 <https://doi.org/10.1016/j.neuroimage.2014.03.018>

Dado Thirza, Güçlütürk Yağmur, Ambrogioni Luca, Ras Gabriëlle, Bosch Sander, van Gerven Marcel, Güçlü Umut (2022) **Hyperrealistic Neural Decoding for Reconstructing Faces from fMRI Activations via the GAN Latent Space** *Scientific Reports* **12** <https://doi.org/10.1038/s41598-021-03938-w>

Défossez Alexandre, Caucheteux Charlotte, Rapin Jérémy, Kabeli Ori, King Jean-Rémi (2023) **Decoding Speech Perception from Non-Invasive Brain Recordings** *Nature Machine Intelligence* **5**:1097–1107 <https://doi.org/10.1038/s42256-023-00714-5>



- Diedrichsen Jörn, Kriegeskorte Nikolaus (2017) **Representational Models: A Common Framework for Understanding Encoding, Pattern-Component, and Representational-Similarity Analysis** *PLOS Computational Biology* **13** <https://doi.org/10.1371/journal.pcbi.1005508>
- Dosenbach Nico U. F. *et al.* (2017) **Real-Time Motion Analytics during Brain MRI Improve Data Quality and Reduce Costs** *NeuroImage* **161**:80–93 <https://doi.org/10.1016/j.neuroimage.2017.08.025>
- Friston K. J., Glaser D. E., Henson R. N. A., Kiebel S., Phillips C., Ashburner J. (2002) **Classical and Bayesian Inference in Neuroimaging: Applications** *NeuroImage* **16**:484–512 <https://doi.org/10.1006/nimg.2002.1091>
- Friston K. J., Holmes A. P., Worsley K. J., Poline J. P., Frith C. D., Frackowiak R. S. J. (1994) **Statistical Parametric Maps in Functional Imaging: A General Linear Approach** *Human Brain Mapping* **2**:189–210 <https://doi.org/10.1002/hbm.460020402>
- Gaziv Guy, Bely Roman, Granot Niv, Hoogi Assaf, Strappini Francesca, Golan Tal, Irani Michal (2022) **Self-Supervised Natural Image Reconstruction and Large-Scale Semantic Classification from Brain Activity** *NeuroImage* **254** <https://doi.org/10.1016/j.neuroimage.2022.119121>
- Giordano Bruno L., Esposito Michele, Valente Giancarlo, Formisano Elia (2023) **Intermediate Acoustic-to-Semantic Representations Link Behavioral and Neural Responses to Natural Sounds** *Nature Neuroscience* :1–9 <https://doi.org/10.1038/s41593-023-01285-9>
- Glover G. H., Li T. Q., Ress D. (2000) **Image-Based Method for Retrospective Correction of Physiological Motion Effects in fMRI: RETROICOR** *Magnetic Resonance in Medicine* **44**:162–67
- Glover Gary H (1999) **Deconvolution of Impulse Response in Event-Related BOLD fMRI** *NeuroImage* **9**:416–29 <https://doi.org/10.1006/nimg.1998.0419>
- Griffin D., Lim Jae (1983) **Signal Estimation from Modified Short-Time Fourier Transform** *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 8. Boston, MASS, USA: Institute of Electrical and Electronics Engineers* :804–7
- Güçlü Umut, Thielen Jordy, Hanke Michael, van Gerven M. A. J., van Gerven Marcel A. J. (2016) **Brains on Beats** *Proceedings of the International Conference on Neural Information Processing Systems* :2101–9
- Gutierrez Miren (2021) **Algorithmic Gender Bias and Audiovisual Data: A Research Agenda** *International Journal of Communication* **15**:439–61
- Han Kuan, Wen Haiguang, Shi Junxing, Lu Kun-Han, Zhang Yizhen, Fu Di, Liu Zhongming (2019) **Variational Autoencoder: An Unsupervised Model for Encoding and Decoding fMRI Activity in Visual Cortex** *NeuroImage* **198**:125–36 <https://doi.org/10.1016/j.neuroimage.2019.05.039>
- Harris Charles R. *et al.* (2020) **Array Programming with NumPy** *Nature* **585**:357–62 <https://doi.org/10.1038/s41586-020-2649-2>

- Higgins Irina, Chang Le, Langston Victoria, Hassabis Demis, Summerfield Christopher, Tsao Doris, Botvinick Matthew (2021) **Unsupervised Deep Learning Identifies Semantic Disentanglement in Single Inferotemporal Face Patch Neurons** *Nature Communications* **12** <https://doi.org/10.1038/s41467-021-26751-5>
- Horikawa Tomoyasu, Kamitani Yukiyasu (2017) **Generic Decoding of Seen and Imagined Objects Using Hierarchical Visual Features** *Nature Communications* **8** <https://doi.org/10.1038/ncomms15037>
- Howard Andrew G., Zhu Menglong, Chen Bo, Kalenichenko Dmitry, Wang Weijun, Weyand Tobias, Andreetto Marco, Adam Hartwig (2017) **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications** *arXiv*
- Kasper Lars *et al.* (2017) **The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data** *Journal of Neuroscience Methods* **276**:56–72 <https://doi.org/10.1016/j.jneumeth.2016.10.019>
- Kell Alexander J. E., Yamins Daniel L. K., Shook Erica N., Norman-Haignere Sam V., McDermott Josh H. (2018) **A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy** *Neuron* **98**:630–644 <https://doi.org/10.1016/j.neuron.2018.03.044>
- Khaligh-Razavi Seyed-Mahdi, Kriegeskorte Nikolaus (2014) **Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation** *PLoS Computational Biology* **10** <https://doi.org/10.1371/journal.pcbi.1003915>
- Kingma Diederik P., Welling Max (2014) **Auto-Encoding Variational Bayes** *arXiv* **1312**
- Kingma Diederik P., Welling Max (2019) **An Introduction to Variational Autoencoders** *Foundations and Trends® in Machine Learning* **12**:307–92 <https://doi.org/10.1561/22000000056>
- Kriegeskorte Nikolaus (2008) **Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience** *Frontiers in Systems Neuroscience* <https://doi.org/10.3389/neuro.06.004.2008>
- Kriegstein Katharina V., Giraud Anne-Lise (2004) **Distinct Functional Substrates along the Right Superior Temporal Sulcus for the Processing of Voices** *NeuroImage* **22**:948–55 <https://doi.org/10.1016/j.neuroimage.2004.02.020>
- von Kriegstein Katharina, Eger Evelyn, Kleinschmidt Andreas, Giraud Anne Lise (2003) **Modulation of Neural Responses to Speech by Directing Attention to Voices or Verbal Content** *Cognitive Brain Research* **17**:48–55 [https://doi.org/10.1016/S0926-6410\(03\)00079-X](https://doi.org/10.1016/S0926-6410(03)00079-X)
- Lavan Nadine (2023) **The Time Course of Person Perception From Voices: A Behavioral Study** *Psychological Science* **34**:771–83 <https://doi.org/10.1177/09567976231161565>
- Le Lynn, Ambrogioni Luca, Seeliger Katja, Güçlütürk Yağmur, van Gerven Marcel, Güçlü Umut (2022) **Brain2Pix: Fully Convolutional Naturalistic Video Frame Reconstruction from Brain Activity** *Frontiers in Neuroscience* **16**
- LeCun Yann, Bengio Yoshua, Hinton Geoffrey (2015) **Deep Learning** *Nature* **521**:436–44 <https://doi.org/10.1038/nature14539>

- Liu Dong, Li Yue, Lin Jianping, Li Houqiang, Wu Feng (2020) **Deep Learning-Based Video Coding: A Review and a Case Study** *ACM Computing Surveys* **53**:11–11 <https://doi.org/10.1145/3368405>
- Maguinness Corrina, Roswandowitz Claudia, von Kriegstein Katharina (2018) **Understanding the Mechanisms of Familiar Voice-Identity Recognition in the Human Brain** *Neuropsychologia* **116**:179–93 <https://doi.org/10.1016/j.neuropsychologia.2018.03.039>
- Maris Eric, Oostenveld Robert (2007) **Nonparametric Statistical Testing of EEG-and MEG-Data** *Journal of Neuroscience Methods* **164**:177–90 <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Martin F. N., Champlin C. A. (2000) **Reconsidering the Limits of Normal Hearing** *Journal of the American Academy of Audiology* **11**:64–66
- Mcallister Jan, Potts Anne, Mason Kathryn, Marchant Geoffrey (1994) **Word Duration in Monologue and Dialogue Speech** *Language and Speech* **37**:393–405 <https://doi.org/10.1177/002383099403700404>
- Millet Juliette, Caucheteux Charlotte, Orhan Pierre, Boubenec Yves, Gramfort Alexandre, Dunbar Ewan, Pallier Christophe, King Jean-Remi (2022) **Toward a Realistic Model of Speech Processing in the Brain with Self-Supervised Learning** *arXiv* **2206** <https://doi.org/10.48550/arXiv.2206.01685>
- Mozafari Milad, Reddy Leila, VanRullen Rufin (2020) **Reconstructing Natural Scenes from fMRI Patterns Using BigBiGAN** *2020 International Joint Conference on Neural Networks (IJCNN)* :1–8 <https://doi.org/10.1109/IJCNN48605.2020.9206960>
- Nagrani Arsha, Chung Joon Son, Zisserman Andrew (2017) **VoxCeleb: A Large-Scale Speaker Identification Dataset** *Interspeech* :2616–20
- Naselaris Thomas, Kay Kendrick N., Nishimoto Shinji, Gallant Jack L. (2011) **Encoding and Decoding in fMRI** *NeuroImage* **56**:400–410 <https://doi.org/10.1016/j.neuroimage.2010.07.073>
- Pasley Brian N., David Stephen V., Mesgarani Nima, Flinker Adeen, Shamma Shihab A., Crone Nathan E., Knight Robert T., Chang Edward F. (2012) **Reconstructing Speech from Human Auditory Cortex** *PLOS Biology* **10** <https://doi.org/10.1371/journal.pbio.1001251>
- Paszke Adam *et al.* (2019) **PyTorch: An Imperative Style, High-Performance Deep Learning Library** *Advances in Neural Information Processing Systems* **32** (NeurIPS 2019)
- Pedregosa Fabian *et al.* (2018) **Scikit-Learn: Machine Learning in Python** *Journal of Machine Learning Research*
- Penhune V. B., Zatorre R. J., MacDonald J. D., Evans A. C. (1996) **Interhemispheric Anatomical Differences in Human Primary Auditory Cortex: Probabilistic Mapping and Volume Measurement from Magnetic Resonance Scans** *Cerebral Cortex* **6**:661–72 <https://doi.org/10.1093/cercor/6.5.661>
- Pernet Cyril R. *et al.* (2015) **The Human Voice Areas: Spatial Organization and Inter-Individual Variability in Temporal and Extra-Temporal Cortices** *NeuroImage* **119**:164–74 <https://doi.org/10.1016/j.neuroimage.2015.06.050>

Petkov Christopher I., Kayser Christoph, Steudel Thomas, Whittingstall Kevin, Augath Mark, Logothetis Nikos K. (2008) **A Voice Region in the Monkey Brain** *Nature Neuroscience* **11**:367–74 <https://doi.org/10.1038/nn2043>

Roswadowitz Claudia, Kappes Claudia, Obrig Hellmuth, von Kriegstein Katharina (2018) **Obligatory and Facultative Brain Regions for Voice-Identity Recognition** *Brain* **141**:234–47 <https://doi.org/10.1093/brain/awx313>

Rupp Kyle, Hect Jasmine L., Remick Madison, Ghuman Avniel, Chandrasekaran Bharath, Holt Lori L., Abel Taylor J. (2022) **Neural Responses in Human Superior Temporal Cortex Support Coding of Voice Representations** *PLOS Biology* **20** <https://doi.org/10.1371/journal.pbio.3001675>

Santoro Roberta, Moerel Michelle, De Martino Federico, Valente Giancarlo, Ugurbil Kamil, Yacoub Essa, Formisano Elia (2017) **Reconstructing the Spectrotemporal Modulations of Real-Life Sounds from fMRI Response Patterns** *Proceedings of the National Academy of Sciences* **114**:4799–4804 <https://doi.org/10.1073/pnas.1617622114>

Schrimpf Martin, Blank Idan Asher, Tuckute Greta, Kauf Carina, Hosseini Eghbal A., Kanwisher Nancy, Tenenbaum Joshua B., Fedorenko Evelina (2021) **The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing** *Proceedings of the National Academy of Sciences* **118** <https://doi.org/10.1073/pnas.2105646118>

Schrimpf Martin *et al.* (2018) **Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?** *bioRxiv* <https://doi.org/10.1101/407007>

Schütt Heiko H., Kipnis Alexander D., Diedrichsen Jörn, Kriegeskorte Nikolaus (2021) **Statistical Inference on Representational Geometries** *arXiv*

Schweinberger S. R., Herholz A., Sommer W. (1997) **Recognizing Famous Voices: Influence of Stimulus Duration and Different Types of Retrieval Cues** *Journal of Speech, Language, and Hearing Research: JSLHR* **40**:453–63 <https://doi.org/10.1044/jslhr.4002.453>

Smith Stephen M. *et al.* (2004) **Advances in Functional and Structural MR Image Analysis and Implementation as FSL** *NeuroImage* **23**:S208–19 <https://doi.org/10.1016/j.neuroimage.2004.07.051>

Snoek Lukas, van der Miesen Maite M., Beemsterboer Tinka, van der Leij Andries, Eigenhuis Annemarie, Steven Scholte H. (2021) **The Amsterdam Open MRI Collection, a Set of Multimodal MRI Datasets for Individual Difference Analyses** *Scientific Data* **8** <https://doi.org/10.1038/s41597-021-00870-6>

Trapeau Régis, Thoret Etienne, Belin Pascal (2022) **The Temporal Voice Areas Are Not ‘Just’ Speech Areas** *Frontiers in Neuroscience* **16** <https://doi.org/10.3389/fnins.2022.1075288>

Tuckute Greta, Feather Jenelle, Boebinger Dana, McDermott Josh H. (2023) **Many but Not All Deep Neural Network Audio Models Capture Brain Responses and Exhibit Correspondence between Model Stages and Brain Regions** *PLOS Biology* **21** <https://doi.org/10.1371/journal.pbio.3002366>

Vallat Raphael (2018) **Pingouin: Statistics in Python** *Journal of Open Source Software* **3** <https://doi.org/10.21105/joss.01026>

- VanRullen Rufin, Reddy Leila (2019) **Reconstructing Faces from fMRI Patterns Using Deep Generative Neural Networks** *Communications Biology* **2** <https://doi.org/10.1038/s42003-019-0438-y>
- Walther Alexander, Nili Hamed, Ejaz Naveed, Alink Arjen, Kriegeskorte Nikolaus, Diedrichsen Jörn (2016) **Reliability of Dissimilarity Measures for Multi-Voxel Pattern Analysis** *NeuroImage* **137**:188–200 <https://doi.org/10.1016/j.neuroimage.2015.12.012>
- Wang Xiaosha, Xu Yangwen, Wang Yuwei, Zeng Yi, Zhang Jiakai, Ling Zhenhua, Bi Yanchao (2018) **Representational Similarity Analysis Reveals Task-Dependent Semantic Influence of the Visual Word Form Area** *Scientific Reports* **8** <https://doi.org/10.1038/s41598-018-21062-0>
- Wetzel Sebastian J (2017) **Unsupervised Learning of Phase Transitions: From Principal Component Analysis to Variational Autoencoders** *Physical Review E* **96** <https://doi.org/10.1103/PhysRevE.96.022140>
- Woods Kevin J. P., Siegel Max H., Traer James, McDermott Josh H. (2017) **Headphone Screening to Facilitate Web-Based Auditory Experiments** *Attention, Perception, & Psychophysics* **79**:2064–72 <https://doi.org/10.3758/s13414-017-1361-2>
- Wu Michael C. K., David Stephen V., Gallant Jack L. (2006) **COMPLETE FUNCTIONAL CHARACTERIZATION OF SENSORY NEURONS BY SYSTEM IDENTIFICATION** *Annual Review of Neuroscience* **29**:477–505 <https://doi.org/10.1146/annurev.neuro.29.051605.113024>
- Xing Chao, Wang Dong, Liu Chao, Lin Yiye (2015) **Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation** *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics* :1006–11
- Yamins Daniel L. K., DiCarlo James J. (2016) **Using Goal-Driven Deep Learning Models to Understand Sensory Cortex** *Nature Neuroscience* **19**:356–65 <https://doi.org/10.1038/nn.4244>
- Zäske Romi, Perlich Marie-Christin, Schweinberger Stefan R. (2016) **To Hear or Not to Hear: Voice Processing under Visual Load** *Attention, Perception, & Psychophysics* **78**:1488–95 <https://doi.org/10.3758/s13414-016-1119-2>
- Zhang Yang, Ding Yue, Huang Juan, Zhou Wenjing, Ling Zhipei, Hong Bo, Wang Xiaoqin (2021) **Hierarchical Cortical Networks of ‘Voice Patches’ for Processing Voices in Human Brain** *Proceedings of the National Academy of Sciences* **118** <https://doi.org/10.1073/pnas.2113887118>

## Editors

Reviewing Editor

**Andrea Martin**

Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Senior Editor

**Barbara Shinn-Cunningham**

Carnegie Mellon University, Pittsburgh, United States of America



**Reviewer #1 (Public Review):****Summary:**

In this study, the authors trained a variational autoencoder (VAE) to create a high-dimensional "voice latent space" (VLS) using extensive voice samples, and analyzed how this space corresponds to brain activity through fMRI studies focusing on the temporal voice areas (TVAs). Their analyses included encoding and decoding techniques, as well as representational similarity analysis (RSA), which showed that the VLS could effectively map onto and predict brain activity patterns, allowing for the reconstruction of voice stimuli that preserve key aspects of speaker identity.

**Strengths:**

This paper is well-written and easy to follow. Most of the methods and results were clearly described. The authors combined a variety of analytical methods in neuroimaging studies, including encoding, decoding, and RSA. In addition to commonly used DNN encoding analysis, the authors performed DNN decoding and resynthesized the stimuli using VAE decoders. Furthermore, in addition to machine learning classifiers, the authors also included human behavioral tests to evaluate the reconstruction performance.

**Weaknesses:**

This manuscript presents a variational autoencoder (VAE) to evaluate voice identity representations from brain recordings. However, the study's scope is limited by testing only one model, leaving unclear how generalizable or impactful the findings are. The preservation of identity-related information in the voice latent space (VLS) is expected, given the VAE model's design to reconstruct original vocal stimuli. Nonetheless, the study lacks a deeper investigation into what specific aspects of auditory coding these latent dimensions represent. The results in Figure 1c-e merely tested a very limited set of speech features. Moreover, there is no analysis of how these features and the whole VAE model perform in standard speech tasks like speech recognition or phoneme recognition. It is not clear what kind of computations the VAE model presented in this work is capable of. Inclusion of comparisons with state-of-the-art unsupervised or self-supervised speech models known for their alignment with auditory cortical responses, such as Wav2Vec2, HuBERT, and Whisper, would strengthen the validation of the VAE model and provide insights into its relative capabilities and limitations.

The claim that the VLS outperforms a linear model (LIN) in decoding tasks does not significantly advance our understanding of the underlying brain representations. Given the complexity of auditory processing, it is unsurprising that a nonlinear model would outperform a simpler linear counterpart. The study could be improved by incorporating a comparative analysis with alternative models that differ in architecture, computational strategies, or training methods. Such comparisons could elucidate specific features or capabilities of the VLS, offering a more nuanced understanding of its effectiveness and the computational principles it embodies. This approach would allow the authors to test specific hypotheses about how different aspects of the model contribute to its performance, providing a clearer picture of the shared coding in VLS and the brain.

The manuscript overlooks some crucial alternative explanations for the discriminant representation of vocal identity. For instance, the discriminant representation of vocal identity can be either a higher-level abstract representation or a lower-level coding of pitch height. Prior studies using fMRI and ECoG have identified both types of representation within the superior temporal gyrus (STG) (e.g., Tang et al., Science 2017; Feng et al., NeuroImage 2021). Additionally, the methodology does not clarify whether the stimuli from different speakers contained identical speech content. If the speech content varied across speakers, the

approach of averaging trials to obtain a mean vector for each speaker-the "identity-based analysis"-may not adequately control for confounding acoustic-phonetic features. Notably, the principal component 2 (PC2) in Figure 1b appears to correlate with absolute pitch height, suggesting that some aspects of the model's effectiveness might be attributed to simpler acoustic properties rather than complex identity-specific information.

Methodologically, there are issues that warrant attention. In characterizing the autoencoder latent space, the authors initialized logistic regression classifiers 100 times and calculated the t-statistics using degrees of freedom (df) of 99. Given that logistic regression is a convex optimization problem typically converging to a global optimum, these multiple initializations of the classifier were likely not entirely independent. Consequently, the reported degrees of freedom and the effect size estimates might not accurately reflect the true variability and independence of the classifier outcomes. A more careful evaluation of these aspects is necessary to ensure the statistical robustness of the results.

<https://doi.org/10.7554/eLife.98047.1.sa2>

### **Reviewer #2 (Public Review):**

#### **Summary:**

Lamothe et al. collected fMRI responses to many voice stimuli in 3 subjects. The authors trained two different autoencoders on voice audio samples and predicted latent space embeddings from the fMRI responses, allowing the voice spectrograms to be reconstructed. The degree to which reconstructions from different auditory ROIs correctly represented speaker identity, gender, or age was assessed by machine classification and human listener evaluations. Complementing this, the representational content was also assessed using representational similarity analysis. The results broadly concur with the notion that temporal voice areas are sensitive to different types of categorical voice information.

#### **Strengths:**

The single-subject approach that allows thousands of responses to unique stimuli to be recorded and analyzed is powerful. The idea of using this approach to probe cortical voice representations is strong and the experiment is technically solid.

#### **Weaknesses:**

The paper could benefit from more discussion of the assumptions behind the reconstruction analyses and the conclusions it allows. The authors write that reconstruction of a stimulus from brain responses represents 'a robust test of the adequacy of models of brain activity' (L138). I concur that stimulus reconstruction is useful for evaluating the nature of representations, but the notion that they can test the adequacy of the specific autoencoder presented here as a model of brain activity should be discussed at more length. Natural sounds are correlated in many feature dimensions and can therefore be summarized in several ways, and similar information can be read out from different model representations. Models trained to reconstruct natural stimuli can exploit many correlated features and it is quite possible that very different models based on different features can be used for similar reconstructions. Reconstructability does not by itself imply that the model is an accurate brain model. Non-linear networks trained on natural stimuli are arguably not tested in the same rigorous manner as models built to explicitly account for computations (they can generate predictions and experiments can be designed to test those predictions). While it is true that there is increasing evidence that neural network embeddings can predict brain data well, it is still a matter of debate whether good predictability by itself qualifies DNNs as 'plausible computational models for investigating brain processes' (L72). This concern is

amplified in the context of decoding and naturalistic stimuli where many correlated features can be represented in many ways. It is unclear how much the results hinge on the specificities of the specific autoencoder architectures used. For instance, it would be useful to know the motivations for why the specific VAE used here should constitute a good model for probing neural voice representations.

Relatedly, it is not clear how VAEs as generative models are motivated as computational models of voice representations in the brain. The task of voice areas in the brain is not to generate voice stimuli but to discriminate and extract information. The task of reconstructing an input spectrogram is perhaps useful for probing information content, but discriminative models, e.g., trained on the task of discriminating voices, would seem more obvious candidates. Why not include discriminatively trained models for comparison?

The autoencoder learns a mapping from latent space to well-formed voice spectrograms. Regularized regression then learns a mapping between this latent space and activity space. All reconstructions might sound 'natural', which simply means that the autoencoder works. It would be good to have a stronger test of how close the reconstructions are to the original stimulus. For instance, is the reconstruction the closest stimulus to the original in latent space coordinates out of using the experimental stimuli, or where does it rank? How do small changes in beta amplitudes impact the reconstruction? The effective dimensionality of the activity space could be estimated, e.g. by PCA of the voice samples' contrast maps, and it could then be estimated how the main directions in the activity space map to differences in latent space. It would be good to get a better grasp of the granularity of information that can be decoded/ reconstructed.

What can we make of the apparent trend that LIN is higher than VLS for identity classification (at least VLS does not outperform LIN)? A general argument of the paper seems to be that VLS is a better model of voice representations compared to LIN as a 'control' model. Then we would expect VLS to perform better on identity classification. The age and gender of a voice can likely be classified from many acoustic features that may not require dedicated voice processing.

The RDM results reported are significant only for some subjects and in some ROIs. This presumably means that results are not significant in the other subjects. Yet, the authors assert general conclusions (e.g. the VLS better explains RDM in TVA than LIN). An assumption typically made in single-subject studies (with large amounts of data in individual subjects) is that the effects observed and reported in papers are robust in individual subjects. More than one subject is usually included to hint that this is the case. This is an intriguing approach. However, reports of effects that are statistically significant in some subjects and some ROIs are difficult to interpret. This, in my view, runs contrary to the logic and leverage of the single-subject approach. Reporting results that are only significant in 1 out of 3 subjects and inferring general conclusions from this seems less convincing.

The first main finding is stated as being that '128 dimensions are sufficient to explain a sizeable portion of the brain activity' (L379). What qualifies this? From my understanding, only models of that dimensionality were tested. They explain a sizeable portion of brain activity, but it is difficult to follow what 'sizeable' is without baseline models that estimate a prediction floor and ceiling. For instance, would autoencoders that reconstruct any spectrogram (not just voice) also predict a sizeable portion of the measured activity? What happens to reconstruction results as the dimensionality is varied?

A second main finding is stated as being that the 'VLS outperforms the LIN space' (L381). It seems correct that the VAE yields more natural-sounding reconstructions, but this is a technical feature of the chosen autoencoding approach. That the VLS yields a 'more brain-like representational space' I assume refers to the RDM results where the RDM correlations were mainly significant in one subject. For classification, the performance of features from the

reconstructions (age/ gender/ identity) gives results that seem more mixed, and it seems difficult to draw a general conclusion about the VLS being better. It is not clear that this general claim is well supported.

It is not clear why the RDM was not formed based on the 'stimulus GLM' betas. The 'identity GLM' is already biased towards identity and it would be stronger to show associations at the stimulus level.

Multiple comparisons were performed across ROIs, models, subjects, and features in the classification analyses, but it is not clear how correction for these multiple comparisons was implemented in the statistical tests on classification accuracies.

Risks of overfitting and bias are a recurrent challenge in stimulus reconstruction with fMRI. It would be good with more control analyses to ensure that this was not the case. For instance, how were the repeated test stimuli presented? Were they intermingled with the other stimuli used for training or presented in separate runs? If intermingled, then the training and test data would have been preprocessed together, which could compromise the test set. The reconstructions could be performed on responses from independent runs, preprocessed separately, as a control. This should include all preprocessing, for instance, estimating stimulus/identity GLMs on separately processed run pairs rather than across all runs. Also, it would be good to avoid detrending before GLM denoising (or at least testing its effects) as these can interact.

<https://doi.org/10.7554/eLife.98047.1.sa1>

### **Reviewer #3 (Public Review):**

#### **Summary:**

In this manuscript, Lamothe et al. sought to identify the neural substrates of voice identity in the human brain by correlating fMRI recordings with the latent space of a variational autoencoder (VAE) trained on voice spectrograms. They used encoding and decoding models, and showed that the "voice" latent space (VLS) of the VAE performs, in general, (slightly) better than a linear autoencoder's latent space. Additionally, they showed dissociations in the encoding of voice identity across the temporal voice areas.

#### **Strengths:**

- The geometry of the neural representations of voice identity has not been studied so far. Previous studies on the content of speech and faces in vision suggest that such geometry could exist. This study demonstrates this point systematically, leveraging a specifically trained variational autoencoder.
- The size of the voice dataset and the length of the fMRI recordings ensure that the findings are robust.

#### **Weaknesses:**

- Overall, the VLS is often only marginally better than the linear model across analysis, raising the question of whether the observed performance improvements are due to the higher number of parameters trained in the VAE, rather than the non-linearity itself. A fair comparison would necessitate that the number of parameters be maintained consistently across both models, at least as an additional verification step.
- The encoding and RSM results are quite different. This is unexpected, as similar embedding geometries between the VLS and the brain activations should be reflected by higher correlation values of the encoding model.

- The consistency across participants is not particularly high, for instance, S1 seemed to have demonstrated excellent performances, while S2 showed poor performance.
- An important control analysis would be to compare the decoding results with those obtained by a decoder operating directly on the latent spaces, in order to further highlight the interest of the non-linear transformations of the decoder model. Currently, it is unclear whether the non-linearity of the decoder improves the decoding performance, considering the poor resemblance between the VLS and brain-reconstructed spectrograms.

<https://doi.org/10.7554/eLife.98047.1.sa0>

#### Author response:

Please find below our provisional author response, outlining the revisions we plan to undertake to address the Recommendations received:

##### **Reviewer #1 (Recommendations For The Authors):**

*(1) A set of recent advances have shown that embeddings of unsupervised/self-supervised speech models aligned to auditory responses to speech in the temporal cortex (e.g. Wav2Vec2: Millet et al. NeurIPS 2022; HuBERT: Li et al. Nat Neurosci 2023; Whisper: Goldstein et al. bioRxiv 2023). These models are known to preserve a variety of speech information (phonetics, linguistic information, emotions, speaker identity, etc) and perform well in a variety of downstream tasks. These other models should be evaluated or at least discussed in the study.*

We plan to evaluate two of these other models, Wav2Vec2 and HuBERT, in the brain encoding and RSA parts.

*(2) The test statistics of the results in Fig 1c-e need to be revised. Given that logistic regression is a convex optimization problem typically converging to a global optimum, these multiple initializations of the classifier were likely not entirely independent. Consequently, the reported degrees of freedom and the effect size estimates might not accurately reflect the true variability and independence of the classifier outcomes. A more careful evaluation of these aspects is necessary to ensure the statistical robustness of the results.*

We plan to address this point to ensure the statistical robustness of our results.

*(3) In Line 198, the authors discuss the number of dimensions used in their models. To provide a comprehensive comparison, it would be informative to include direct decoding results from the original spectrograms alongside those from the VLS and LIN models. Given the vast diversity in vocal speech characteristics, it is plausible that the speaker identities might correlate with specific speech-related features also represented in both the auditory cortex and the VLS. Therefore, a clearer understanding of the original distribution of voice identities in the untransformed auditory space would be beneficial. This addition would help ascertain the extent to which transformations applied by the VLS or LIN models might be capturing or obscuring relevant auditory information.*

We plan to include direct decoding results from the original spectrograms in addition from the VLS and LIN models.



## Reviewer #2 (Recommendations For The Authors):

We plan to address the following points raised by Reviewer #2:

(1) English mistakes, rewordings:

- a. L31: 'in voice' > consider rewording (from a voice?).
- b. L33: consider splitting sentence (after interactions).
- c. L39: 'brain' after parentheses.
- d. L45: certainly DNNs 'as a powerful tool' extend to audio (not just image and video) beyond their use in brain models.
- e. L52: listened to / heard.
- f. L63: use second/s consistently.
- g. L64: the reference to Figure 5D is maybe a bit confusing here in the introduction.
- h. L79-88: this section is formulated in a way that is too detailed for the introduction text (confusing to read). Consider a more general introduction to the VLS concept here and the details of this study later.
- i. L99: again, I think the experimental details are best saved for later. It's good to provide a feel for the analysis pipeline here, but some of the details provided (number of averages, denoising, preprocessing), are anyway too unspecific to allow the reader to fully follow the analysis.

We will correct the mistakes, apply the suggested rewordings, and clarify the points raised.

(2) Clarification.

- L159: what was the motivation for classifying age as a 2-class classification problem? Rather than more classes or continuous prediction? How did you choose the age split?
- L263: Is the test of RDM correlation > 0 corrected for multiple comparisons across ROIs, subjects, and models?
- L379: 'these stimuli' - weren't the experimental stimuli different from those used to train the VAE?
- L443: what are 'technical issues' that prevented subject 3 from participating in 48 runs??
- L444: participants were instructed to 'stay in the scanner'!? Do you mean 'stay still', or something?
- L463: Hearing thresholds of 15 dB: do you mean that all had thresholds lower than 15 dB at all frequencies and at all repeated audiogram measurements?
- L472: were the 4 category levels balanced across the dataset (in number of occurrences of each category combination)?

- L482: the test stimuli were selected as having high energy by the amplitude envelope. It is unclear what this means (how is the envelope extracted, what feature of it is used to measure 'high energy'?)
- L500 was the audio filtered to account for the transfer function of the Sennheiser headphones?
- L500: what does 'comfortable level' correspond to and was it set per session (i.e. did it vary across sessions)?
- L526- does the normalization imply that the reconstructed spectrograms are normalized? Were the reconstructions then scaled to undo the normalization before inversion?
- L606: does the identity GLM model the denoised betas from the first GLM or simply the BOLD data? The text indicates the latter, but I suspect the former.
- L704: could you unpack this a bit more? It is not easy to see why you specify the summing in the objective. Shouldn't this just be the ridge objective for a given voxel/ROI? Then you could just state it in matrix notation.
- L716: you used robust scaling for the classifications in latent space but haven't mentioned scaling here. Are we to assume that the same applies?
- L720: Pearson correlation as a performance metric and its variance will depend on the choice of test/train split sizes. Can you show that the results generalize beyond your specific choices? Maybe the report explained variance as well to get a better idea of performance.
- Could you specify (somewhere) the stimulus timing in a run? ISI and stimulus duration are mentioned in different places, but it would be nice to have a summary of the temporal structure of runs.

We will clarify the points raised.

### **Reviewer #3 (Recommendations For The Authors):**

We plan to address the following points raised by Reviewer #3:

#### **Comments:**

- Code and data are not currently available.
- In the supplementary material, it would be beneficial to present the different analyses as boxplots, as in the main text, but with the ROIs in the left and right hemispheres separated, to better show potential hemispheric effect. Although this information is available in the Supplementary Tables, it is currently quite tedious to access it.
- In Figure 3a, it might be beneficial to order the identities by age for each gender in order to more clearly illustrate the structure of the RDMs,

- *In Figure 3b, the variance for the correlations for the aTVA is higher than in other regions, why?*
- *Please make sure that all acronyms are defined, and that they are redefined in the figure legends.*
- *Gender and age are primarily encoded by different brain regions (Figure 5, pTVA vs aTVA). How does this finding compare with existing literature?*

We will upload the code and the preprocessed data; improve the supplementary material figures; Fix Figure 3 according to the Reviewer's suggestion, and clarify the points raised.

<https://doi.org/10.7554/eLife.98047.1.sa4>